DOCUMENT RESUME

ED 218 075	SE 037 440
AUTHOR	Cole, Henry P.; Moss, Judson
TITLE	Evaluation of Learning for Continuing Engineering Education. Draft Report, Revised.
INSTITUTION	Kentucky Univ., Lexington. Coll. of Engineering.
SPONS AGENCY	National Science Foundation, Washington, D.C.
PUB DATE	Jan 80
GRANT	NSF/SED-80008; SED-78-22060
NOTE	394p.
EDRS PRICE	MF01/PC16 Plus Postage.
DESCRIPTORS	*Academic Achievement; *Course Evaluation; Curriculum Development; Curriculum Evaluation; Engineering; *Engineering Education; Evaluation Methods; Formative Evaluation; Higher Education; Measures (Individuals); Pretests Posttests; *Professional Continuing Education; Science Education; Student Characteristics; Summative Evaluation; *Testing
IDENTIFIERS	National Science Foundation

ABSTRACT

This report documents the measurement of learning outcomes and the formative/summative evaluation of courses as derived from the activities and experiences of educators in many fields of endeavor to determine how best to design courses of instruction and how to measure the learning outcomes resulting from courses. The report is divided into four major parts. The first part describes a general typology of continuing education courses, the characteristics of persons enrolled in such courses, and the use of formative and summative evaluation in course development. The second part is a detailed explanation concerning the use of various types of testing procedures in measuring learning outcomes. The third part describes alternative methods for developing valid and reliable tests to measure individual learning achievement and overall course effectiveness. Methods for reporting the results of learning assessments to individuals and groups are presented in the last section. The report is written such that parts of it may be useful to individuals with specific needs without having to read the whole text and is facilitated by a detailed table of contents and a detailed subject index. A precis of the 14 chapters is included in the preface. (Author/JN)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

NSF/SED-80008

SE078-22060

ED218075

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

✓ This document has been reproduced as received from the person or organization originating it.
Minor changes have been made to improve reproduction quality.

- Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

DRAFT REPORT

EVALUATION OF LEARNING FOR CONTINUING ENGINEERING EDUCATION

By

Henry P. Cole & Judson Moss

University of Kentucky
Lexington, Kentucky

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

National
Science Foundation

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

SE 037 440

Evaluation of Learning
for
Continuing Engineering Education

by
Henry P. Cole, Ed.D.[1]
assisted by
Judson Moss, Ed.D.[2]
and

Steering Committee, Measurement for Learning Outcomes
in Continuing Education for Scientists and Engineers
Project Members: Billy J. Barfield, Ph.D.,
Professor David K. Blythe, Frank Gohs, M.S.,
Edward Kifer, Ph.D., Warren Lacefield, M. S.,
Donna Mertens, Ph.D., and Joanne Wilson, M. S.
(University of Kentucky, Lexington, KY)

[1] Professor and Chairman, Department of Educational Psychology College of Education, University of Kentucky, Lexington, KY.

[2] Project Manager, Learning Outcomes Measurement Project, Office of Continuing Education, College of Engineering, University of Kentucky, Lexington, KY.

Evaluation of Learning
for
Continuing Engineering Education

Measurement for Learning Outcomes
in Continuing Education for Scientists and Engineers
Office of Continuing Education,
College of Engineering, University of Kentucky
Lexington, KY

October 1979

Revised January 1980

These Guidelines were prepared with the support of the National Science Foundation, Grant No. SED78-22060. Any opinions, findings, conclusions, or recommendations expressed herein are those of the authors and do not necessarily reflect the views of the National Science Foundation.

PREFACE

This book is written particularly for the director of continuing education in engineering and related technical fields. The book has developed out of the activities of a group of persons with diverse talents and backgrounds, all of whom were involved in a project concerned with the measurement of learning outcomes for continuing education courses in engineering. Members of this group consisted of a highly experienced director of continuing education in engineering; professors of continuing education courses in engineering; experienced specialists in adult and higher education with much experience in developing and teaching continuing education courses in a variety of technical fields including the health professions; and educational psychologists expert and highly experienced in the areas of measurement of human abilities and skills, the design of tests, and educational program and course evaluation.

For a period of two years this group has met on a regular basis in an ongoing seminar about the measurement of learning outcomes for a variety of courses typical of those offered in continuing education programs for engineers at many colleges and universities and in other settings as well. In addition, members of the group have worked together as small teams in the actual development and use of methods and procedures for the measurement of the learning outcomes resulting from a number of continuing education courses in engineering taught under the direction of the University of Kentucky, College of Engineering, Office of Continuing Education and Extension.

The book is a set of detailed guidelines which bring together what the project team has learned about how to go about the measurement of learning outcomes in this field. Much of what is presented in the book concerning the measurement of learning outcomes and the formative and summative evaluation of courses has been derived from the activities and experiences of many other educators in many fields in efforts to determine how to best design courses of instruction and how to measure the learning outcomes resulting from courses. What is new is the bringing of all of this information together in the context of adult education and specifically in the area of engineering courses designed for continuing education purposes. Consequently, the way in which specific procedures bear on the measurement of learning outcomes in these courses is well illustrated in many examples.

There are four main sections in this book. The first part describes a general typology of continuing education courses, the characteristics of persons enrolled in such courses, and the use of formative and summative evaluation in course development. The second part of the book is a detailed explanation concerning the use of various types of testing procedures in measuring learning outcomes. A third section describes alternative methods for developing valid and reliable tests to measure individual learning achievement and overall course effectiveness. A fourth section presents methods for reporting the results of learning assessments to individuals and groups. The reader is advised to scan the table of contents, read Chapters 1 and 14, and then to select those sections of the text of most interest.

The book is written such that parts of it may be useful to individuals with specific needs without their having to read the whole text. This objective is further facilitated by the very detailed table of contents and a detailed subject index. There is information about a wide range of topics which should be of value to directors of continuing education in engineering as well as to the professionals who develop and teach such courses.

Some chapters are more geared to the specific procedures concerning how good courses may be developed and how their learning outcomes may be measured in efficient and yet effective ways. The information presented in these chapters has relevance to the design and evaluation of any course, although the examples presented are specifically in engineering and in the continuing education context. Chapters 4, 5, 7, and 10 can be used as the basis for study by instructors who have an interest in improving the effectiveness of their course in reaching intended learning outcomes.

Other chapters address matters of great interest to the administrators and policy makers who operate and oversee continuing education programs. Chapter 2 provides a classification of four basic typologies of continuing education courses and some insight into how the typology of the course effects both its delivery and its evaluation. Chapter 3 reminds the reader of the salient differences between continuing education courses and the students typically enrolled in them and the more traditional undergraduate and graduate formal courses which are part of college and university degree programs. Implications for the staffing of such courses, their scheduling, and their evaluation are noted.

Chapters 4 through 9 describe in great detail the various types of tests and testing procedures which are available and useful to the business of designing courses and measuring their effectiveness in terms of the achievement of learners on a variety of performance measures. All the procedures are based upon stating in operational terms the intended learning outcomes related to the performance of persons in the work setting.

Chapter 10 outlines a very detailed but general and useful set of procedures for insuring that well organized courses and valid and reliable measurement procedures are developed. Chapter 11 describes empirical ways of determining the degree of validity and reliability of tests, test items, and other learning assessment procedures which have been developed for purposes of making inferences about the degree of group and individual learning resulting from a course.

Chapter 12 presents important limitations of tests as assessment devices. This chapter is important because persons should not misuse test data in the construction of inferences about the degree of success of individual students and course effectiveness.

Chapter 13 presents detailed procedures and information about how to report the data gathered from course evaluations and individual student achievement. How this information should be used, with whom it should be shared, in what manner, and for what purposes are all discussed.

Finally, Chapter 14 is a summary of the entire book in that a set of recommendations are made for the development of an "evaluated CEU." All the strengths and limitations of the procedures available for the measurement of individual learning by students and for judging the effectiveness of courses are recalled. This information is used to conclude with a recommendation that courses and programs be evaluated and certified rather than individuals.

All of these chapters may be of interest and value to the continuing educator charged with being accountable to professional agencies, individuals, and administrative superiors for the quality and effectiveness of the program and courses operated under his or her jurisdiction.

The book has proven to be of interest to directors and faculty in continuing education in other technical fields. such as nursing and the allied health professions. Although all of the examples provided are in continuing engineering education, what is presented is expected to be generally useful to continuing education activities in many areas.

ACKNOWLEDGEMENTS

Although this book was drafted by primarily one author, it is truly a group effort. All of the persons listed on the title page made continuing and important contributions. The members of the project committee met over a two year period for regular seminars during which much of the specific content of the book and many of the topics were generated, discussed at length, and often debated in a lively manner. In addition, individual members of the project group were assigned to the development of evaluation instruments and procedures and to the use of these procedures in the evaluation of the learning outcomes of a number of continuing education courses. The courses are part of a program offered under the supervision of Associate Dean David K. Blythe, Director of Continuing Education, College of Engineering, University of Kentucky.

Dean Blythe's experience and knowledge of continuing education and engineering were invaluable. He spent many hours educating the members of the group in key matters, directing the primary author to many sources which are cited in the referenced in the book, and providing a tutorial in many aspects of the topic with which the book deals. In addition, Dean Blythe was responsible for bringing the group together and for obtaining the funds from the National Science Foundation which made the project possible.

Dr. Judson Moss, the Project Manager and an experienced specialist in adult and continuing education, also provided many of these same types of services. He along with Dean Blythe prepared the proposals which funded the project activities. He also contributed much specialized knowledge and expertise in the area of adult education and wrote the portions of Chapter 3 which deal with adult learner characteristics. In addition he coordinated and organized all the multiple activities of the project members, compiled all of the data from the studies completed, and collected, edited, wrote, and rewrote all of the project reports with the assistance of other project members. In addition, Dr. Moss continually directed the primary author of the book to important works and studies in continuing education in engineering and in other related fields. Furthermore, Dr. Moss edited and proofread all of the chapters in the book as they were drafted by the primary author. His cheerful and very capable assistance in this task and in the much more complex task of managing the project is greatly appreciated and was essential to the completion of the project activities and the completion of the book.

Mr. Warren Lacefield, Mr. Frank Gohs, and Dr. Edward Kifer along with the primary author developed most of the measurement scales, questionnaires, and other instruments used in the actual evaluation of the courses studied in the project. In addition, these persons did the bulk of all of the data processing and analysis which allowed the project staff and the course instructors to determine how effective the courses were and how valid and reliable the test and measurement procedures. Many of the examples presented in the book include the instruments and the data developed by these persons. Mr. Gohs and Mr. Lacefield also wrote Appendix A. In addition, Mr. Gohs wrote some of the material included in Chapter 13, particularly about the Urban Storm Water Quality Modeling course used as an example of how to measure and report learning outcomes.

Dr. Donna M. Mertens designed, conducted, and analyzed the results of the evaluation of the Engineering Economics course which was included as one of the courses studied. In addition she wrote the initial draft of the last section of Chapter 3 concerned with the motivations of professionals for attending continuing education courses.

Dr. Billy Barfield and his partner Dr. Tom Haan, authors of Hydrology and Sedimentology of Surface Mined Lands (1978) provided a sterling example of a well designed and operated continuing education course. Many of the examples in the book are based upon practices of these professors in the teaching of this well-organized and excellent course. In addition, Dr. Barfield participated in the ongoing seminars and discussions and provided the invaluable insights and criticisms of an experienced engineer and continuing education professor.

Ms. Joanne Wilson, Mr. Frank Gohs, and Dr. Edward Kifer all provided the primary author a continuing and stimulating dialogue about the topics of the book in another ongoing doctoral seminar in educational program and product evaluation. This seminar was taught in the Department of Educational Psychology and Counseling during the fall 1979 semester. All three individuals attended regularly, were part of the project team, and greatly influenced the content of the book as it was drafted by the primary author. In addition, all three persons, as have other members of the project team, were critical readers of the early drafts of the work and made many suggestions leading to its improvement. Mr. Warren Lacefield did an exceptionally detailed critical reading of the second draft of the manuscript and made many improvements in both the style and content of each chapter.

Ms. Melanie Barber typed the several drafts of each chapter, accurately, efficiently, and with good humor. She also provided assistance in checking and locating references.

Many professors of continuing education courses in engineering made the project possible by letting teams of project members come into their classes and participate in the instruction, submitting to interviews and many questions, and developing tests for their courses, modifying the tests to improve them after subsequent analysis, and seeing to it that the tests were administered to the persons enrolled in their courses. The author and the project staff are particularly indebted to Dr. Billy Barfield, Dr. Thomas Haan, and Dr. Michael Meadows.

It has been an honor to be part of this group and to have been extended the privilege of being the primary spokesperson for the collective wisdom of the staff.

Henry P. Cole

636 Bellcastle Road
Lexington, Kentucky 40505

March 30, 1980

TABLE OF CONTENTS

Chapter Title Page

1 INTRODUCTION AND OVERVIEW 1

Part 1 - Basic Properties of Continuing Education Courses, Their Development, and Evaluation

2 CHARACTERISTICS OF CONTINUING ENGINEERING EDUCATION COURSES 12

Four Basic Types of Courses 12
Remediation 13
Extending Prior Knowledge and Skill 13
Imparting Advanced Technical Concepts and Skills 14
Exposure to Knowledge Outside of Engineering Science 15
Different Instructional Purposes and Methods Across Course Types 16
Assessing Learning Across Course Types 19
Utility of Learning Outcome Assessments 21
Limitations of Typical Learning Assessment Procedures 23
Conclusion 25

3 CHARACTERISTICS OF COURSE PARTICIPANTS 27

Focus on Practical Needs 27
Professional Engineers as Students 31
Adult Learners: Andragogy Versus Pedagogy . . . 35
Roles of Adult Learners in the Learning Process 36
Roles of Adult Learners in Evaluation of Courses 38
Motivations for Attending Continuing Education Courses 41
Conclusion 46

4 FORMATIVE EVALUATION AND COURSE DEVELOPMENT 49

Stages of Course Development 49
Selecting Course Content 52
Refining Course Objectives and Learning Assessment Tasks 55
Measuring Only That Which is Appropriate . . . 60
Packaging and Delivery Considerations 65
Selection of Instructors 69
Specialized Equipment and Facilities: Implications for Course Delivery and Learning Assessment 72
Practitioners' Tacit Evaluation of Courses . . 74
Conclusion 75

Chapter Title Page

5 SUMMATIVE EVALUATION OF LEARNING OUTCOMES 77

Need for Summative Evaluation. 78
Differences Between Summative and Formative Evaluation 80
Four General Learning Assessment Procedures . . 81
Developing Sound Learning Assessment Procedures 81

Part 2 - Using Tests to Measure Learning Outcomes

6 PRE-TESTS, THEIR PURPOSES AND USES 84

Pre-tests as Informative and Screening Devices 84
Pre-tests as Devices for Adjusting Course Content and Operation 87
Pre-tests as Indicators of Baseline Performance 88
Learning Resulting from Pre-test Experiences. . 93
Two Approaches to Constructing Pre-tests . . . 95

7 EMBEDDED TESTS: THEIR PURPOSES AND USES 101

Traditional Embedded Tests 101
Abbreviated Forms of Embedded Tests 106
An Example of a Course with Embedded Tests . . 107
Reasonable Expectations for Achievement Within a Short Course 112
Practical Problems in Using Complex Embedded Test Tasks in Short Courses 113
Abbreviated Embedded Test Tasks: An Illustration 115
Purposes and Properties of Abbreviated Embedded Test Tasks 117
Advantages of Multiple Choice Items as Embedded Test Tasks 1. 120
Precautions in Developing Multiple Choice and Other Objective Test Items 121
Importance of Item Independence 124
Generalization of Item Construction Procedures to Other Test Formats 126
Generalization of Item Construction Procedures to Pre-, Post, and Delayed Post Test Construction 128

Conclusion 130

Chapter Title Page

Calculation of Item Difficulty and Discrimination Indices for Criterion Referenced Tests 213
Item Analysis Procedures in Perspective . . . 215
Methods of Reliability Estimation: The NR and CR Cases 216
Alternate Forms Method 217
Test Re-Test Method 225
Sub-divided Test Method 228
Internal Consistency Methods 228
Conclusion 231

12 LIMITATIONS OF TESTS 235

Sources of Invalidity and Unreliability . . . 235
The Standard Error of Estimate 240
Stability of Group Mean Scores 243
Advantages of Criterion Referenced Tests . . . 246
Summary of Major Points Concerning Testing . . 254

Part 4 - Sharing Learning Assessments and Course Evaluations

13 REPORTING THE ASSESSMENTS OF LEARNING OUTCOMES 259

Participants' Needs 259
Instructors' Needs 260
Program Administration Needs 262
Client Agency Needs 263
Professional Societies' Needs 263
Meeting Diverse Information Needs 264
Basic Information: Student Achievement, Course, and Instructor Characteristics . . . 266
Gathering and Presenting Basic Achievement Data: An Example 269
Reporting Learning Outcomes to Individual Students 282
Keeping Learning Outcomes of Individuals Private 294
Precautions to Prevent the Abuse of Test Scores 298
Making Course Evaluations Public 303
Conclusion 305

14 CONCLUSION - RECOMMENDATIONS FOR EVALUATED CEUs 306

Courses Where Tests Provide Accurate Estimates of Learning 306
Courses Where Tests are Inadequate Estimates of Learning 308
Limited Time for Testing 309
Inadequacy of Testing in Sampling the Performance Domain 311

Chapter Title Page

Growth of Learning After Course Completion . . 312
The Need for Multiple Indicators of Learning Outcomes 313
The Impossibility of Making "Complete" Learning Assessments of Individuals 315
Means for Making Comprehensive Assessment of Course Effectiveness 316
Logical Requirements for Certification of Courses 318
Logical Requirements for Certification of Persons 318
The Importance of Options for Participants . . 320
Involving All Participants in Learning Assessment Activities 321
Conclusion 324

REFERENCES 325

APPENDIX A 331

APPENDIX B 344

AUTHOR INDEX 368

SUBJECT INDEX (incomplete)

LIST OF TABLES

Table Page

1 Engineers' Interest in Non-Formal Education Programs 32

2 Rank Ordering of Motivational Factoss for Participating in Continuing Education . . . 44

3 Steps for Developing Pre-tests, Embedded Tests, Post Tests, and Delayed Post Tests, by Which to Estimate Learning Outcomes 161

4 Performance Objectives for Open Channel Hydraulic Structures Unit: An Illustration of Test Construction Procedures . . . . . . 167 & 348

5 Test for "Open Channel Hydraulics" Unit - Illustrating the Mapping of Items to Performance Objectives . . . . 170 & 350

6 Scores of Engineers on Alternate Forms of Three Criterion Referenced Mastery Tests . . 220

7 Pre- and Post Test Items and Their Assignment to Test Forms for the Urban Storm Water Course 272

8 Pre-test Total Scores Across Test Forms - Urban Storm Water Quality Modeling Course . 274

9 Post Test Total Scores Across Test Forms - Urban Storm Water Quality Modeling Course . 276

A-1 Demographic Information Questionnaire 333

A-2 Participant Reaction Questionnaire 336

A-3 Satisfaction/Utilization Survey Form 340

A-4 Structured Personnel Interview Protocol . . . 342

10 Limiting Velocities and Inactive Forces for Open Channels 353

11 List of Considerations for Preparing Multiple Choice Items 366

LIST OF FIGURES

Figure Page

1 Pre- and Post Test Scores by Persons with Test Means, Standard Deviations, and Persons by Score Regression Lines for a Short Course 89

2 Illustration of a Graphic Means for Reporting Learning Outcome Measurements for a Short Course to Individual Participants and Groups 278

3 Learning Outcomes Resulting from a Short Course for Engineers on Urban Storm Water Quality Modeling 279

4 Learning Outcomes Resulting from a Short Course for Engineers on Urban Storm Water Quality Modeling 280

5 Sample Standard Answer Sheet for Manual or Machine Scoring 286

6 Manual Individual Achievement Reporting Form . . 289

7 Computerized Individual Achievement Reporting Form 293

8 Properties of Typical Channels 353

9 n-VR for Various Retardance Classes 354

10 Solution for Manning's Equation, Vegetated Waterways 355

11 Solution for Manning's Equation, Vegetated Waterways 356

Chapter 1

INTRODUCTION AND OVERVIEW

This book is intended as a set of practical guidelines for directors of continuing engineering education and others involved in the task of helping to update the knowledge, skills, and practice of the Nation's many engineers. The guidelines may also be of value to those persons involved in the continuing education of other scientific and technical specialties.

The need to involve engineers and other technical professions in continuing education activities throughout their careers is grounded in a number of factors. First, the present rate of knowledge expansion and technology insures that continuing education in a broad range of scientific and technical topics is required for maintaining competence among engineers and other technical personnel. Professional and scientific journals assist toward this end as do formal courses of instruction and programs in the engineering sciences at colleges and universities.

In addition, many industrial organizations engage in research and development and provide instruction for their technical staffs in the new knowledge developed from their own and others' activities. Yet, there remains a strong need for the systematic organization and efficient presentation of basic and newer technical knowledge to a wide audience of engineers through short courses, conferences, workshops, evening classes, and similar activities (Klus & Jones, 1975). Many of the Nation's engineers work in small organizations not able to mount the ongoing technical training programs common to some of the larger firms. In addition, even the largest engineering firm or company is not capable of offering the wide number of continuing education courses and delivery modes needed even by their own employees, much less meet the needs of technical personnel from other agencies and areas. Well organized and managed continuing education programs for the engineering sciences need to be consistently available. A wide range of courses needed by different persons is required. In addition, multiple modes of course delivery are needed.

For example, practicing engineers today need courses in areas as diverse as human relations skills, engineering economics, recent technical developments in micro-processing equipment, and effects of specific environmental toxins and their proper management. Engineers increasingly have become involved in long-term community and state planning. Engineers are frequently the coordinators of industrial and community development groups, the persons responsible for knowing and insuring compliance with environmental protection laws, and major consumers and users of recent technical developments. It is impossible to teach the breadth of knowledge and skills required to satisy this range of assignments in one or two professional degree programs completed at a university or college. It is also unrealistic to expect that this wide range of skills and knowledge will automatically develop simply through "on the job training." Each of these areas contain large amounts of technical information and specific skills which often require additional systematic instruction beyond that which can be received in preparatory professional programs at colleges and universities. Acquisition of this additional technical information by the engineer, and his or her greater facility in specific technical skills, are expected learning outcomes from most continuing education courses. It is the measurement of these and other related types of learning outcomes with which this book is concerned.

The best ways to assist engineers in the acquisition of these types of knowledge and skills depend on a number of peripheral factors. These include: a) the geographic location and distribution of the persons needing the particular type of instruction; b) the content of the course, its complexity, and its optimal duration; c) the specific learning outcomes sought as the result of the course: e.g., increased awareness of the law or available technology, specific improved performance in technical areas such as use of micro-processors in the operation and use of industrial production processes; and d) the characteristics of the engineering students who will be involved: e.g., prior relevant technical training, previous work experience, recency of formal courses and technical work in areas such as mathematics and computer programming.

For all of these reasons, once a continuing education need in a specific technical area is identified, one cannot simply produce a standard course taught by traditional means. Decisions concerning whether to offer the course as a short course at a three day conference, as a once-a-week evening class, as a correspondence course, or as a mail-out TV cassette tape course (with accompanying readings, workbooks, and manuals) should be based on considerations of geographic distribution of participants, availability of qualified and competent instructional personnel, the size of the prospective student group, and many other factors. Other decisions concerning course content and level(s) of difficulty, the rate of presentation of material, the optimum duration of instruction, and the amount of prior participant skill and knowledge also have to be made. These decisions depend heavily upon the characteristics of the engineers to be enrolled.

In short, to be effective, continuing education programs must not only teach content and skills needed by practicing engineers, but in doing so must adapt to the characteristics and limitations of the persons needing the instruction. This makes offering a sound continuing engineering education program a difficult task requiring much wisdom and good information about the needs, characteristics, and activities of these practicing professionals in their work roles.

These same factors also complicate the evaluation of learning outcomes for continuing engineering education courses. The decisions about the level(s) of difficulty of courses, of duration and mode of delivery, and of the major intended outcomes, all directly influence the methods and procedures used to evaluate course effectiveness and the degree of individual student learning. The design and implementation of the instructional system itself should not be separated from the evaluation activities used to assess degree of individual student learning and judge overall course effectiveness. This principle was noted long ago by Ralph Tyler (1950), and has consistently been observed to be basic to good practice in the area of educational course and program evaluation (Bellack & Kliebard, 1977). This principle should be clear in order that the reader be disabused of the idea that there is one simple "best" method or set of procedures by which to evaluate the outcomes of continuing education courses in engineering. There is neither one "best" method or procedure for the teaching or evaluation of such courses! There are, however, some well-established guidelines and alternative strategies for constructing and evaluating continuing education courses in technical fields with attention given to the variables noted earlier.

In large part the present level of expertise in program evaluation has grown out of concerns about the effectiveness of large federally-supported curriculum development activities (Grobman, 1968; Worthen & Sanders, 1973). These and earlier curriculum development activities in public and higher education, as well as in military training activities, provide a good foundation for approaching the topic of "evaluation of learning outcomes" in continuing engineering education. It is the purpose of this book to set forth this accumulated knowledge in the hope that it will serve as a set of useful procedures for engineering educators. The procedures offered are also grounded in the activities of an interdisciplinary group of scholars fron engineering, higher education, and educational psychology. This group has been engaged in the evaluation of learning outcomes for multiple and diverse continuing education courses in the engineering sciences.[1]

The concepts of both formative and summative evaluation will be used throughout this book (Bloom, Hastings, & Madaus, 1971; Greenfield, 1978; Grobman, 1968). It should be clear to the reader that these terms are being applied to the class of courses concerned with the upgrading of knowledge and skills of practicing engineers and related technical personnel through a variety of short courses and professional training seminars and workshops. The guidelines presented are not intended for the development of formal courses for undergraduate or graduate engineering courses taught over the course of a semester in typical college or university programs which lead to a degree. However, many of the principles set forth are useful in these situations as well.

---

[1] The activities of this group were supported by the National Science Foundation, Grant No. SED78-22060, The Learning Outcomes Measurement Project.

The focus is upon alternative methods by which to evaluate the learning outcomes for an array of short courses designed for the professional engineer. It is necessary to attend to the purposes, objectives, content, organization, packaging, delivery, and follow-up activities which are usually part of such short courses if they are to be evaluated properly. Evaluation is an integral part of the development and delivery of any such course (Aleamoni, 1980; Gagne, 1967; Grobman, 1968). This type of evaluative activity is usually called formative evaluation, the means by which courses are improved toward more effectively meeting the needs of the persons who enroll in them.

Summative evaluation is also required in order to report to the individual participant something about the degree to which he or she has achieved the course objectives, what more needs to be learned, and how competently the procedures and skills learned in the course may be practiced in daily professional activity upon returning to the work site. Summative evaluation is also needed to describe the general effectiveness of such courses in order that businesses and agencies who send engineers and other technical personnel to participate in training can be informed about the general effectiveness of the training. It also provides information required for periodic adjustments and improvements in continuing education courses and programs by the persons who develop and operate them. Summative evaluations are descriptive and judgmental statements about the intentions, procedures, and worth of courses or programs taken as whole units. They literally provide a summary of course operation and effectiveness in meeting desired learning outcomes over a given time period with specific groups of enrollees. Any summative evaluation may also be used in a formative way to make improvements in future replications of courses and their operation within programs.

The remainder of this book is organized into four parts. The chapters in the first part describe a typology of continuing education courses. Different types of courses serve different purposes and must be evaluated in different ways. Another chapter deals with the characteristics and needs of persons typically enrolled in continuing education courses. Two additional chapters deal with formative and summative evaluation and describe how these activities can be used to help courses and programs to meet the needs of the participants who enroll in them.

The second part of the book provides a detailed description about the use of four different types of tests in the measurement of learning outcomes resulting from instruction. The information and examples provided illustrate how the various types of tests should be used, when they should be used, and how the tests should be developed.

A third part of the book describes alternative procedures for the development of valid and reliable tests and measuring instruments by which to make inferences about the degree of individual learning and overall course effectiveness. The development of tests and the testing activity is presented as an integral part of the process of instructional development activities which are required to create high quality instruction. The instructional purposes and uses of tests and testing are emphasized. One chapter is devoted to specific methods of test item analysis and test reliability determination. Another chapter is devoted to the limitations of tests and testing and serves to place this method of assessment of individuals' learning and evaluation of course effectiveness in a proper perspective.

The fourth part of the book is a single, but extensive chapter which describes and provides examples of how assessments of individuals' learning and evaluations of course effectiveness can and should be reported to various persons and groups. The needs of individual students, instructors, administrators, client agencies, and professional societies for information about the effectiveness of courses in achieving their intended learning outcomes are described. Methods of meeting these diverse needs for information about the performance of courses and the persons who teach and complete these courses are described. The types of information needed, methods for gathering the information, and effective ways of presenting the resulting data are all described. The privacy of individuals' test scores and other learning assessments and the public nature of the overall evaluation of course and program effectiveness are emphasized. Precautions are suggested to prevent the abuse of test scores.

Means for the evaluation of "courses and programs" rather than "persons" are presented in the last chapter. Courses where testing provides accurate estimates of learning are described along with other courses where this is not the case. Limited time for testing, problems of adequate sampling of the performance domain being instructed, and the anticipated growth in learning of course content after course completion are all presented as problems which need to be recognized and dealt with in the measurement of learning outcomes resulting from instruction. The impossibility of making "complete" assessments of an individual's learning resulting from a given course is noted. The use of multiple indicators of learning outcomes, item sampling procedures, and comprehensive assessments of course effectiveness are recommended rather than only the assessment of individual persons' knowledge and skill by common testing procedures. The last chapter pulls together the previous sections of the book and makes strong recommendations. It may be useful to read the last chapter first and then proceed to other sections and chapters of the book as it suits the needs and interests of the reader.

The detailed table of contents and this first chapter should be of assistance to persons in making the decision about where to begin. Although the book is written in a sequential manner, it is also intended that persons having interest in any particular topic or section will be able to readily locate that section and profit from its study without having to read other sections of the text.

Chapter 2

CHARACTERISTICS OF CONTINUING ENGINEERING EDUCATION COURSES

There is a variety of common types of short courses for engineers. Courses differ in purposes, content, delivery method, packaging, and intended audience. Each of these distinguishing characteristics will be examined and implications for evaluation of the different varieties of learning outcomes noted.

Four Basic Types of Courses

Continuing education courses for professional engineers may be classified into four broad categories. These include courses concerned with:

a) remediation and upgrading of basic knowledge and skills;

b) extending and broadening previously learned scientific and technical concepts and skills;

c) imparting new concepts and skills in technical areas at high levels of expertise to keep pace with advancing scientific discovery and technology;

d) acquiring new awareness, knowledge, and concepts outside the areas of engineering and basic sciences in order to perform in a more effective manner the many other duties required of engineers in the areas of economics, personnel and business management, ecology and environmental protection, and community development, organization, and problem solving.

Let us now consider examples of each of the four types of courses. Frequent reference will be made to this typology of continuing education courses throughout this book.

Remediation

The most common courses in this category are short courses designed to improve basic concepts and skills to help engineers pass state licensing examinations. These courses are popular with recent graduates of engineering programs who plan to sit for such examinations.

Another example includes short courses designed to sharpen skills and concepts once learned, thought not recently used, but now in demand because of a new process being widely adopted. Basic electrostatics and bonding concepts in physics and chemistry, which were once learned by most engineers but since forgotten, are an example of an area in need in a period when many firms are developing copying processes similar to xerography.

Extending Prior Knowledge and Skill

General courses which relate sophisticated concepts and principles in different areas to one another for purposes of integrating and extending concepts learned earlier are found frequently in this category. An example might be a course on casting of metal alloys from powdered metal in wet environments. This course might involve information from dentistry, industrial engineering research concerning fabrication of certain machine parts, development of new materials for repair of structural cracks, and low temperature castings. Generally, the persons enrolled in the course are seeking more information about current theory, research, and technology about a given topic across some array of applications. Such courses emphasize updating and integrating the prior knowledge of the participants with recent developments in related fields. Frequently the motivation for attending such a course is to increase one's general knowledge about the topic area and keep abreast of recent developments. The expected learning outcome anticipated by the participants and proposed by the course developers is often not a skill in relation to improved on-the-job performance. Rather, it is a more general knowledge and awareness of relationships.

It is worth noting that some participants may also enroll in other types of courses for this same "general knowledge improvement" outcome, even though the course may be targeted to teach specific skills and concepts very related to job performance in specific technical areas. This means that the typology of a given course is determined not only by the course objectives and content, but by the intentions, expectations, and motivations of course participants as well. More will be said about this in later chapters.

Imparting Advanced Technical Concepts and Skills

Courses on microcircuits and microprocessors to update the engineer on the latest thinking and development in this field and to call attention to the many potential applications to his or her area of work, are one example of this type of course. Typically such courses are highly specialized and focused on specific learning outcomes. They often require high levels of expertise and knowledge in prerequisite capabilities. For example an advanced course concerning the use of microprocessors to aid industrial production control processes would usually require participants to be facile in computer programming, electronics, optimal control theory, and the basic mathematical procedures which underlie all three areas. The participants enrolled in these types of advanced technical courses often are motivated to attend because they expect to learn very specific methods and procedures directly relevant to their ongoing work.

Exposure to Knowledge Outside of Engineering Science

Courses on current environmental standards for air and water pollution designed to inform engineers of the allowable limits, methods of analysis, and methods to determine cost efficient ways to reduce levels of chemicals, particles, and other by-products and waste materials in the environment are examples in this category. Courses concerned with teaching the skills of human relations and communications to engineers in order that the village or city civil engineer can learn how to mediate more effectively between opposing and often highly antagonistic individuals from business, industry, union, consumer, and environmental protection groups are still another example. The need for courses similar to these examples increasingly is being recognized. The city, village, or project engineer must often mediate among diverse community groups and get them to work together in the task of meeting environmental standards for air and water purity or plan long term community development. Most engineers have had little formal training in the skills required for effectively assuming this role. Yet, engineers are frequently called upon to serve in facilitating, counseling, and group leadership roles. Because of this deficit in their prior professional education, continuing education needs exist for courses in engineering and community economics, in communications and leadership skills, and in conflict resolution and human relations skills to better enable engineers to be more effectively involved in community development activities.

Different Instructional Purposes and Methods Across Course Types

It should be noted that the different types of courses have not only different intended outcomes, but that the means by which to measure these outcomes are also different. Courses of the first type, concerned with remediation of basic knowledge and skill for purposes of scoring well on a professional licensing exam, are similar to typical college courses. For the course concerned with basic knowledge and skills the most appropriate outcome measure is the success of the course participants in passing the professional licensing examination. The most appropriate measures for assessing performance before, during, or after the course are test items sampled from the content domains on which the persons will be tested on the professional licensing examination.

One of the most appropriate learning activities for courses directed toward remediation and upgrading of basic knowledge and skill is the completion of many practice problems sampled from the broad domain on which the student will be tested. These characteristics also have implications concerning the most effective duration, organization, and delivery of the course.

With the basic remediation course, distributed practice with many sessions interspersed with homework, the correction of homework problems and frequent quizzes is the most reasonable approach. These assessment procedures serve to inform the learner and the instructor of an individual's progress. Areas needing additional study by individuals and group instruction by the teacher are clearly identified during the course of instruction. A different approach to teaching and assessment of learning is required for courses in the other areas, such as short duration, intensive seminars or workshops. These other types of courses may be very effective, particularly if the participants have the necessary prior knowledge, skills, and an active need for the new information, but they should not be planned, taught, or evaluated in the same way.

The third type of course, concerned with imparting specialized high levels of technical knowledge and skill to a person already highly skilled in an area, is probably the type of course for which prerequisite knowledge and skill levels are most important. Without the appropriate level of entry skills for such a course, little learning can occur because participants will be unable to understand and perform the learning activities comprising the course. Therefore, in these courses, some sort of screening by a performance task or test, and adequate listing of prerequisite skills for participants before registration is very important. This can often be accomplished by clearly stating the course prerequisite skills and knowledge requirements when advertising the course. Pre-tests may also be used to determine the entry level skill and knowledge of participants, and to advise persons of their readiness for the course. Pre-tests also serve to communicate to prospective particpants useful information about specific course content.

Assessing Learning Outcomes Across Course Types

What can be accepted as evidence of achievement of the intended learning outcomes for these different types of courses varies across the four categories. In most cases, measures which sample the individual's performance on the job or on tasks similar to those encountered on the job are the best indicators of competence. Since it is not possible to include such real performance learning opportunities and assessment tasks within courses in full measure, it is necessary to design learning activities and testing situations which sample some of the key parts of the performance required on the job. A good performance task for an advanced technical course on microprocessors could include the assembly of electronic circuits which would perform an information processing task. The validation of the performance of the circuit and its proper interfacing with laboratory equipment might be another type of performance task by which to assess the degree of learning. The proper operation of the circuit and the student's ability to validate the circuit in a laboratory simulation is a rigorous test of the student's performance. No other test is needed. These types of assessment tasks can be built into the instruction of the short course, and, indeed, are most appropriately administered in this context.

In contrast, in a course designed to foster increased skills in human relations and communications, it may not be possible to so thoroughly assess the acquired learning outcomes. Often the performance measure most appropriate in such situations is the presentation of a series of filmed role plays or verbally presented episodes. These can be followed by paper and pencil tests requiring the analysis of the roles and characteristics played by various persons in the episodes. Based upon course principles participants can be asked to develop a plan or course of action for some real-life sutuation involving the meeting of disparate groups he or she is expecting to work with in the future. An additional requirement might be that the plan be developed as a product the course instructor can evaluate.

For both the microprocessors and human relations skills courses, the best performance measure is follow-up observation of the degree to which the course particpants put the new knowledge and skill to use and the competence they exhibit in doing so. However, such information, beyond self-report or supervisor ratings, is difficult to obtain. For this reason, test tasks and other assessment methods used within the confines of a course, or sent to participants as a follow-up activity and returned for scoring by the course instructors, are often more practical and can be very helpful in assessing the quality of the course and the amount of student retention and learning.

Utility of Learning Outcome Assessments

It should be apparent that the information obtained from the assessment of course learning outcomes is valuable to several groups. First, it is of great value to the persons who develop, teach, and redesign the courses themselves. Any course can be improved if systematic information about how well it reaches selected learning outcomes is attended to. It is often necessary to ask why these outcomes have been realized or not realized, and not only how well. Answering the why question often calls for good descriptions of how the course operated, the attitudes and behavior of the participants, and instructors, the appropriateness of the content, and the physical features of the situation in which the course was instructed. For example, even the best designed course may not work well if offered by a belligerent instructor, late in the evening, in a poorly ventilated and illuminated room. Such conditions may cause participants to lose interest, become hostile, not pay attention, or drop out of the course. Consequently, it is important to systematically seek information about how courses are presented and managed as well as about the learning outcomes of participants on tests and other performance tasks (Worthen & Sanders, 1973). Questionnaires, interviews, observations of classes, and similar methods are often very useful toward this end.

Another group with strong interest in learning outcomes of short courses and related training activities is the participants themselves. Contrary to popular opinion, participants do not mind taking tests and being assessed by other performance measures (Ferry, 1979; Moss, Barfield, & Blythe, 1978). However, the tests or procedures must be reasonable in length and difficulty, directly related to the area of instruction, and the results useful to the learner in self-assessment of present levels of skill or knowledge. The results of assessments of participants' learning should always be promptly reported to the individual. It should also be done privately. An individual's performance assessment should not be made as a public announcement (Wolf, 1974). Later in this book methods and means of efficient reporting of learning outcomes of participants will be described and illustrated.

Employers are another group of persons with legitimate interests in the results of learning outcomes of workers who have participated in continuing engineering education courses, especially since the employer often provides release time from work and payment of tuition costs for the participant. Employers have a right to know the past general success rate of particular courses with groups of participants similar to their own employees. In addition, they have legitimate interests in the progress of individual employees.

Professional societies and the administrative officials who oversee continuing education programs have similar needs.

Limitations of Typical Learning Assessment Procedures

Caution must be observed in reporting results of learning assessments and tests of individual participants as the only measure of or the definition of learning which has occurred. Typical assessment procedures or tests always test less than the functional domain of skill and knowledge. Very important learning outcomes are often not measured by any given test or assessment procedure. For example, a person may complete a course on the construction of sedimentation basins and stream channels for control of water runoff on surface mined lands. Perhaps this person scores very poorly on the post test for the course, indicating he has learned little. However, the person may be a former land surveyor whose present job in state government requires him to have considerable skill and knowledge in this new area, perhaps because he is an inspector. Now suppose this individual is very worried about his lack of knowledge in this specialty. Perhaps he decides to study the course materials on his own after the course. Because of the course and the results of the post test he can now identify what basic knowledge, computational algorithms, and procedures he needs to understand. Perhaps, also because of the course, the individual is able to identify another professional in attendance at the course to whom he can turn for assistance. Suppose the participant seeks the assistance of this colleague in future assignments and that, additionally, he reports his own assessment of his learning needs to his supervisor and requests that he be allowed to attend the course again. If he follows this plan of action, he has, indeed, exhibited some very major and important learning outcomes, despite his low post test score. Certainly his behavior is a likely indication of eventual improved performance in his job. In any event, the individual has achieved an important learning outcome, a more informed knowledge of his present limitations and how to correct them.

One can also anticipate the reverse situation in which a person scored high on the formal assessment procedure or test for the course, but really remained functionally unable or unwilling to put into daily practice what she or he had learned.

The point of all of this is simple. Test scores and other short and artificial assessment procedures are only tentative indications of what a person knows and has learned. Actual performance and change in performance and work activities on the job are much better indicators of the value of the course with respect to learning outcomes for a given course. There is a tendency for professional engineering societies and academics concerned with documentation of amount of learning resulting from continuing education courses and CEUs to ignore these facts.

Conclusion

In summary, many engineers, engineering educators, and some members of professional engineering societies expound the view that there ought to be a straight forward way by which to measure the learning outcomes of courses. The viewpoint is commonly stated as an expectation that it is logical and possible to assess learning in such courses in terms of a given test with a test score or some other numerical performance score. While it is certainly possible to design good tests which measure aspects of learning in very reliable ways, and while it is possible to devise very good functional performance tests, it remains very difficult and costly to make a thorough assessment of the learning outcomes of a particular course for a particular engineer. The common and naive wish for a simple 10 minute pre-test and another 10 minute post test by which to ascertain the precise amount of learning of an individual continuing engineering education student in a given course is not realistic. It is not that such tests cannot be developed to be highly reliable and valid to a specific outcome. They can be quite easily. Rather, it is that the inclusiveness of the test is almost always much less than the intended and important outcomes of the short course. Therefore, when one uses such a test, one knows only that the participants know so much or so little about that specific aspect of the course and its content. It is not appropriate to generalize from that one test score to matters of: a) whether or not the person will use the knowledge gained through the course in actual practice; b) whether he or she can actually use the knowledge or skill in varied and real work related tasks; and c) whether or not the person has learned anything of lasting significance simply because this particular test score is high or low.

Recognizing these limitations should not cause the reader to despair. There are many methods by which reasonable estimates may be made about the effectiveness of specific courses in achieving their intended, and sometimes unintended, learning outcomes. Later chapters in this book detail many of these methods.

Chapter 3

CHARACTERISTICS OF COURSE PARTICIPANTS

The participants enrolled in continuing engineering education courses are distinctly different from persons enrolled in undergraduate and even graduate courses in engineering degree programs at colleges and universities. This is an important point for it reminds us that the persons who voluntarily attend short courses, on their own or as representatives of their companies or agencies, do so for reasons different from those of persons enrolled in courses within degree programs. Consequently, instruction should be different for this population.

Focus on Practical Needs

One difference is that the short course participants are already practicing engineers or engineering technologists, in most cases. The usual concerns that faculty members exhibit about the quality of the engineering student and his or her qualifications as a practitioner are moot points with the short course enrollee. By and large these persons are already practicing engineering, qualified or unqualified. They are qualified by virtue of holding the position for which they earn a living doing engineering within some agency or business. The concern of the short course instructor should be much more on the functional, efficient, safe, and wise performance of the engineer in some special area for which the course is specifically designed.

A second difference between short courses of this type and more traditional college and university courses is that the former are more focused on specific skills, concepts, and procedures, while traditional courses are often much more diverse, general, and theoretical. In traditional courses in engineering and basic sciences, professional engineering societies and academics decide what knowledge and skill is basic to practicing engineering. This core becomes the content of the curriculum. In short courses for practicing engineers, academic standards, basic theory, and curriculum objectives of traditional academic courses are all secondary to the functional needs of the paracticing engineer who is often returning to learn something specific about some area of performance which needs to be improved or some topic of special interest.

Michael Scriven, (1977), an expert in educational program evaluation, makes a very nice distinction between objectives and needs. He notes that when teachers and other operators of social programs have a captive audience, they talk about "meeting the objectives of the program." These are objectives of the person of persons who design, develop, and operate the program. However, when the audience is not a captive one, when persons are spending their own money in search of new ideas, skills, or things, the focus is almost never on "objectives" but on "needs".[2] This distinction is the basic difference between traditional undergraduate and graduate engineering courses in degree programs and short courses of a continuing education nature for professional engineers.

The persons who come to such courses care very little about the objectives of the instructor or the university. They care a great deal about meeting their personal needs related to performing better in their work. The objectives of the course are useful to these engineers only insofar as they communicate to the individual participant the nature of the course and what may be learned from participation in the course. Objectives can also provide information about prerequisite levels of skill prior to entering the course, information useful to the person in assessing his or her own readiness to enter the course and profit from the experience.

Likewise, evaluations of the typical learning outcomes from such courses for earlier groups are of interest to participants and employers. These past evaluations provide information about the anticipated utility of the course prior to spending one's own money and time in completing it.

---

[2] Comments made in an address to the students and faculty of the University of Kentucky, Graduate School, by Michael Scriven, March 12, 1979.

Evaluations of the general effectiveness of such courses are something akin to "fairness in labeling" in medication, consumer products, and other areas. A company or an individual has a right to know the general effectiveness of a course in terms of improved knowledge, skill, and performance on the job. In fact, it is just such types of evaluation, usually accomplished informally by questioning and by observation by co-workers and supervisors of persons who have completed specific short courses, which are the basis for future course enrollment and success.

In short, if the course is well designed to meet the needs of the participants in particular areas of focus, and if there are not readily available and efficient ways to meet this need other than the short course, then the course is likely to be very successful. Its effectiveness will be recognized by the persons functionally engaged in this area of practice. The course subsequently will be praised by word of mouth in professional and corporate circles, and consequently become heavily enrolled. It is good that this practical form of tacit evaluation exists and it ought to be encouraged.

This distinction between persons enrolling in continuing education short courses and more traditional degree program courses also has implications for how students should be instructed. Before detailing these differences in instructional methodology and style, it will be helpful to review information about the typical characteristics of persons enrolled in engineering short courses.

Professional Engineers as Students

Adult learners who participate in continuing engineering education programs differ in a number of respects from undergraduate and graduate students. These differences have implications for both administrators and instructors of such programs. A survey of some 257 continuing engineering education participants at courses offered by the University of Wisconsin pointed out some of the major differences (Klus & Jones, 1975). This survey revealed that continuing education participants are usually: a) older, having a median age of 30-30 years; b) practicing engineers rather than full-time students; and c) seeking to upgrade previous knowledge and skills for direct and immediate, rather than deferred, application to their jobs. In addition to participation in formal college or university course work, fifty-four percent of the practicing engineers in the survey reported strong interest in various types of continuing education programs and activities (See Table 1). The completion of in-house educational programs was ranked highest in terms of a preferred mode for continuing education. The completion of short courses and similar continuing education activities sponsored by professional societies and governmental agencies was also rated highly.

Study through formal college credit courses was least valued as a mode of continuing education.

Table 1

Engineers' Interest in
Non-formal Education Programs
(N = 257)

Type of Educational Program	Percentage of Persons Expressing a Preference for This Type of Educational Program
Reading current engineering and technical literature	54%
Completing minimum 1 credit course	18%
Completing in-hourse educational programs	74%
Completing professional, society, & government sponsored programs	55%

Source: Klus, J. P. & Jones, J. A. Engineers involved in education: A survey analysis. Washington, D.C.: American Society for Engineering Education, 1975.

Because these course participants are usually working full-time in their profession, the time which they have available for instruction and study is more limited than their graduate and undergraduate counterparts. Therefore, instructors are forced to develop and use time-efficient methods in their course offerings for continuing education participants.

While many undergraduate and graduate students in formal college and university programs would like more instruction in practical matters, the typical graduate engineers enrolled in a continuing education course often demand learning experiences with immediate application. In the case of more general courses designed to integrate concepts and principles rather than focus on particular skills, participants demand a sharp focus on specific topics and time efficient instruction. Rarely is more than a few days available for the presentation of the material. Employers have similar concerns.

A recent survey of graduate engineers indicated that 65 percent of the participants in a group of continuing education courses in engineering at the University of Kentucky said an important consideration influencing their attendance was that their expenses were paid by their employers (Mertens, 1979). Employers are very concerned with the direct relevance of courses to the daily work activities of their employees. This fact is widely known to persons operating continuing education courses and interacting regularly with employers of engineers. Given this concern and the fact that many engineers do wish to attend short courses and other forms of continuing education activities, it is important to recognize the strong need for applied and practical courses.

Since most university-based continuing engineering education programs are taught by professors who also teach at the graduate and undergraduate levels, it becomes necessary for these instructors to "shift gears" in order to provide the learning outcomes desired by participants in continuing education courses. Faculty development and inservice education activities may be beneficial to staff from engineering colleges who usually teach regular courses in formal degree programs. These professors often can learn much about the needs of graduate engineers in continuing education courses, the characteristics of these learners which distinguish them from younger undergraduate students, and means to develop and present effective and efficient short courses to meet these needs. The teaching methods and organization of instruction in short courses need to be more focused, skillfully articulated and executed than in courses where more time is available. The course content also needs to be more sharply delineated. Professors accustomed to teaching more traditional college courses often have much to learn in these and related areas before they can become effective developers and teachers of continuing education courses. A comprehensive listing and description of the specific skills and competencies needed by continuing education faculty is provided by McCullough (1980). He also provides a questionnaire for assessing faculty competence in each major area of performance. The questionnaire and the performance categories upon which it is based are useful in determining what specific skills need to be developed in persons who are assigned the task of developing, operating, and evaluating continuing education courses for engineers and other technical personnel.

The preparation of faculty for this "new" role can be facilitated by the involvement of experts from other disciplines such as adult education, educational psychology, and instructional design. Any inservice education of faculty in the design and operation of continuing education courses also ought to include direct experience with well designed continuing education courses and the persons who develop and regularly teach these courses. The most exemplary teachers and courses may sometimes be from outside the academic community of major universities and colleges. Inservice education efforts are frequently needed to assist engineering faculty to meet the challenges posed by a different, more adult, experienced, and practially oriented clientele encountered in short courses. The focus on teaching very specific topics and skills within a very restricted time frame also demands major adjustments in teaching style.

Adult Learners: Andragogy versus Pedagogy

In recent years adult educators have differentiated the assumptions concerning adult learners from those traditionally linked to children. One such educator, Malcolm Knowles (1970), has coined the term "andragogy" to describe the art and science of helping adults learn, as contrasted with "pedagogy", the art and science of teaching children. Knowles' concept is based upon four assumptions concerning changes which occur when a person matures. These include: a) the movement of self-concept from dependency toward self-direction; b) the accumulation of experience which becomes a valuable learning reservoir; c) the orientation of the individual's learning readiness to the developmental tasks of his or her social roles; and d) the shifting of time perspective from postponed to immediate application of learning. As learners become adult their orientation increasingly turns from one of subject-centeredness to problem-centeredness (Knowles, 1970).

The acceptance of these assumptions concerning adult education has implications for administrators and instructors active in continuing engineering education.

Roles of Adult Learners in the Learning Process

If adult learners are self-directed, in varying degrees, it follows that they need to become involved in the total learning process as much as possible. This includes: a) assisting in planning the learning objectives and activities; b) actively engaging in the learning experience itself; and c) evaluating the learning experience in terms of its outcomes and their worth.

The persons who develop and operate continuing education courses for engineers should routinely involve samples of actual persons employed in the target area for their courses. These persons, as well as their employers, should be asked to make judgments and comments about the proposed objectives, content, organization, and teaching location for courses which are under development. Sometimes surveys of the educational needs of the population of engineers in a region should be undertaken before courses are developed. This type of activity is commonly referred to as conducting a "needs assessment". Gathering such information about needs and exposing proposed course objectives, materials, and procedures to prospective enrollees can serve two purposes. First, it can improve the course which is finally developed. Second, it can alert professional engineers and their employers that there is a local group of engineering educators who are competent in many areas and genuinely interested in the needs of working professionals. These activities improve courses and instruction and also build rapport with the professional engineering community, which otherwise may be unaware of the local expertise and resources available to them. These experienced adults who make their daily living by doing engineering activities are very qualified and capable of planning what they need to learn and how to accomplish it.

Adult learners should be directly involved in the teaching-learning activities with each other and with the instructors in ongoing courses. Course participants will frequently have example applications, problems, and prior experiences which can be shared and can amplify the points made by course activities and the instructors. It is important that instructors respect the maturity and experience of participants and recognize the value of seeking out and using their contributions, criticisms, observations, and ideas. Arrogant instructors who feel that participants are often incompetent in basic areas and must be told what to do and must always bow to the expertise of the instructor, will usually run into much difficulty with such groups. The more appropriate relationship is one of a tutor, expert in some areas, who seeks to share some of this expertise with fellow professionals who have actively committed the time and energy to come together and study, question, and learn about an area of interest to them. Some college and university instructors accustomed to playing the authoritative professor role in traditional courses have difficulty adopting the partnership role necessary for continuing education courses.

Roles of Adult Learners in Evaluation of Courses

The adult learners enrolled in continuing education courses are also a very important source of judgment about the effectiveness of the courses in achieving certain outcomes, some expected and others unexpected. This group also can judge the worth of courses, and parts of courses, in promoting worthy outcomes. Worth is usually defined as learning something useful which somehow facilitates one's work activity or some aspect of this activity. Consequently, the judgments of course participants, and their employers and supervisors, about the worth of courses is important information which should be routinely sought out, collected, and processed by continuing education program operators. Formal testing and performance assessment procedures, as well as interviews and questionnaires, also can provide information about the worth of courses and lead to their improvement.

Instructors and administrators charged with the responsibility of continuing engineering education have traditionally shied away from formal evaluation techniques. Such persons often feel that adult learners are reluctant to be judged by their peers and, thus, would perhaps not attend programs providing this type of evaluation of individual student learning outcomes. Recent studies have shown that this is generally a false assumption (Ferry, 1979; Moss et al., 1978). Interviews conducted by the Learning Measurement Project and earlier studies by the University of Kentucky's College of Engineering, reveal that participants are willing to have objective evaluations made of their learning. This is especially so if the persons who teach the courses view the learners as adults and convey the genuine desire to evaluate the course and program and not only the individual learner. Students who recognize that the results of formal testing and other performance assessment procedures will be used to provide evaluations of course effectiveness and lead to improved course organization will generally be quite willing to be assessed. This is even more true if the students are aware that the tests and assessment procedures have been carefully constructed and are reasonable estimates of their knowledge and skill in specific aspects of the course.

While program administrators must be aware that particpants may be reluctant to take tests, particularly if they are poorly designed and improperly used, they should also be aware of the growing demand from many sources for valid vertification of learning outcomes. These sources include employers, who often pay all or part of course fees; universities; professional societies; and accrediting agencies. Although additional research is needed in this area, studies completed to date appear to show that engineers attending continuing education courses do not resist objective measures of performance, particularly if these measures help them assess their own level of learning, and contribute to course improvement. The key question should always be, "How effective is this course?" After that question is answered it is appropriate to ask, "How much did this particular participant learn about specific intended outcomes and how can his or her learning be accurately estimated?"

Adequate evaluation, it must be recognized, can be costly, both in time and money. However, objective evaluation of learning outcomes can serve several purposes. These include: a) meeting the demands for accountability of outside agencies; b) enabling continuing education administrators to more effectively meet the needs of their clientele; c) providing course instructors with objective data to determine how much learning has resulted from instruction; d) providing a method for evaluation of effectiveness of course instructors; and e) demonstrating to course participants and employers objective methods for determining course learning outcomes.

Motivations for Attending Continuing Education Courses

The engineer's motivation for participating in continuing education has implications for both the advertising of programs and the instructional design of courses within continuing education programs. Information about participants' motives for attendance can be used to direct the content of the promotional materials to potential enrollees and their employers, thus resulting in attracting larger numbers of interested students. In addition, course content and instructional techniques can be designed to enhance student motivation and achievement (Cole, 1980).

The general motivation of adults for participating in continuing education has been studied by a number of researchers (Monstain & Smart, 1974). However, literature specifically about the engineer's motivation is scarce. Wiesehugel (1978), using a taxonomy proposed by Miller (1977), studied the motivational factors in a group of professional engineers. Wiesehugel found that the most popular reasons for attending continuing education courses were payoff from previous study, acknowledgement of a changing knowledge base and the need to remain abreast, absence of accepted certification in one's field, and upward aspirations. Wiesehugel's results are of interest because his sample consisted of professional engineers. However, his study was not as conceptually or methodologically sound as would have been desirable.

Based on the work of Wiesehugel and Monstain and Smart, an exploratory study was conducted by Mertens (1979). This study was part of the Learning Outcomes Measurement Project conducted at the University of Kentucky. The nature of the study and a summary of the results are presented below. Although the results are based upon only one course, which falls into the fourth category of course typologies, the results are informative.

The Mertens (1979) study was based on responses of 179 professionals enrolled in a television-presented course in economic analysis for engineers. The course was sponsored by the Appalachian Education Satellite Program, and the participants were located at 25 sites in Appalachia. The particpants represented primarily civil engineers (25%), followed by "other" (22%), mechanical engineers (19%), electrical engineers (15%), and chemical engineers (2%). The "other" group included non-engineers, engineering assistants and technologists, and planners.

The participants completed a pre-course survey which included questions about demographic information, sources of tuition fees, and whether employers had recommended attendance. In addition, 19 items believed to be related to participants' motivation for attending the course were listed. Participants were asked to rate the importance of each item for determining their decision to enroll in the course. A five point Likert scale was used to rate each item.

The results of the participants' response to the motivational factors are presented in Table 2. The mean ratings of the 179 participants are presented in rank order of importance. The highest rated items pertained to professional advancement, e.g., "To learn new ideas that might enhance my job performance", and "To acquire specific knowledge of a field or subject." The lowest rated items were related to external expectations or influences, e.g., "I want the certificate that is awarded at the end of the course," and "My agency/supervisor strongly recommended that I attend." It is also interesting to note that the

Table 2

Rank Ordering of Motivational Factors for Participating in Continuing Education*

	Motivational Factor	Mean Rating	s.d.
1.	To learn new ideas that might enhance my job performance.	1.44	.66
2.	Interest in the subject.	1.46	.62
3.	To acquire specific knowledge of a field or subject.	1.48	.72
4.	To learn some of the newer techniques of economic analysis.	1.80	.89
5.	To acquire a new set of skills.	1.84	.81
6.	The opportunity for professional advancement.	1.97	.96
7.	The opportunity for intellectual stimulation.	2.21	1.16
8.	The location was within commuting distance.	2.25	1.22
9.	To do economic analysis.	2.39	1.11
10.	To refresh my skills in an already familiar area.	2.64	1.27
11.	Have taken continuing education courses prior to this one and found them to be of value.	2.91	1.19
12.	To meet with my colleagues and exchange ideas with them.	3.02	1.09
13.	My boss wants me to go to school	3.19	1.36
14.	Taking this course may help me maintain my present situation.	3.26	1.33
15.	My expenses were paid.	3.28	1.45
16.	My agency/supervisor strongly recommended that I attend.	3.44	1.34
17.	Need more education.	3.44	1.28
18.	Want the certificate that is awarded at the end of the course.	3.49	1.32
19.	Know other engineers who are better off because they took a course in engineering economy.	3.62	1.12

*Rating Scale:
1 - very important
2 - moderately important
3 - neutral
4 - moderately unimportant
5 - very unimportant

standard deviations of the highly rated items are uniformly small, while the standard deviations of the lowly rated items are much greater. This indicates that there was much common agreement among participants about the highly ranked items as motivational factors for attending the course and less agreement about the other factors ranked lowly.

More information is required in order to correctly interpret the ratings of several items. For example, the participants' response to "My expenses were paid," is meaningful only if the particpants' expenses were indeed paid. When asked who paid for their attendance, 63 percent responded that their employers paid for them. The mean rating of this 63 percent of the participants for the item (15), "My expenses were paid", was 2.89. This contrasts with the overall mean rating of 3.28. Thus, those whose expenses were paid rate this factor as having somewhat more importance.

This situation also applies to the item, "My agency/ supervisor strongly recommended that I attend." Forty-five percent of the participants indicated that their employer recommended attendance. The mean rating for this group for the above item (16) was 2.84, as contrasted with the whole group mean rating of 3.44. Once again, the group for whom the item was most relevant rated the item as slightly more important.

Mertens' (1979) results suggest two conclusions: First, engineers enrolled in this course tend to rate the items concerning acquisition of specific knowledge and skills and professional advancement as most important for influencing their participation in continuing education. These are predominently self directed or intrinsic motives. Secondly, external influences appear to be less important for determining participation. However, the rating of these external factors is directly influenced by the relevance of the item for the individual participant. This suggests that the course attracted a variety of different persons for different reasons. This is probably a very common situation for most courses in continuing education areas. Yet, as established by the rank ordering of the items based on participant responses, it is clear that the self-directed learning motives basic to andragogy theory are predominant.

Conclusion

Earlier sections of this chapter have called attention to the concern of many engineers for continuing education courses of a practical nature. The Mertens study as well as common experience suggest this is a strong concern. However, this point should not be over emphasized. A survey of engineers by Morris, Sherrill, and Scriven (1978) indicated that 40 percent of the respondents would establish policies to support a continuing education program in which up to

50 percent of the programs had non-job related content. This finding is not suprising if one recognizes that most engineers are curious persons who have strong and lasting interests in many areas of science and technology, some directly related to their work and others remotely or unrelated (Holland, 1973).

Many engineers enroll in courses specifically designed to teach them knowledge and skills for a particular engineering activity, even when that activity is not in the domain of their work performance. It is common to find a wide variety of persons from other engineering and non-engineering vocations in courses such as "The Hydrology and Sedimentology of Surface Mined Lands," a course designed specifically for civil and mining engineers. A mechanical engineer from an equipment manufacturing company may attend because of curiosity and a desire to understand more clearly the problems his clients face. Even though the engineer is not anticipating any specific objective or outcome which will result from his learning, he will frequently find the course produces beneficial results in the future. The course may help him better understand aspects of the work of mining and civil engineers. It may cause him to read more and add to and broaden his general knowledge.and appreciation of technology and science. Knowledge like money is a very generalizable currency. Once it has been acquired there are almost always exciting, worthwhile, and often surprising, ways to use it. Knowledge is even better than money, because it is not consumed in use. Rather it is strengthened. Perhaps the experienced engineer who is a curious, life long learner, characterized by an andragogy outlook, is more aware of this than most persons.

Continuing education courses in engineering need to be offered in all four categories or typologies. The adult characteristics of the participants need to be recognized. The motives for continuing study should be understood by instructors and participants should be encouraged to enroll in whatever types of courses they need for whatever reasons. These different purposes of courses and the different motives of persons within the same course insure there will be many different learning outcomes for any given course as well as for different types of courses. This confounds the easy assessment of learning outcomes. The intention of the course developers, the intentions of the participants, and the actual operation and teaching - learning methodology of any course all help determine what may or may not be learned from a course. Evaluations of courses and their effectiveness and individual assessments of course participants' knowledge and skill must take these factors into consideration.

Chapter 4

FORMATIVE EVALUATION AND COURSE DEVELOPMENT

Good continuing education engineering courses develop gradually through several stages. Seeking and using appropriate information about the need for such courses, their presentation and effectiveness can lead to the development of a course which teaches participants what they need to know in a consistent and efficient pattern. The collection and analysis of information about the early operation of courses for the purpose of insuring a more effective learning operation in the future is called formative evaluation. It is the process whereby initial course designs to meet the needs of participants.are refined and revised. The business of assessing learning outcomes for improving a particular course requires different types of information at different stages.

Stages of Course Development

Short courses of a continuing education nature often develop in response to some particular need of practicing engineers. This is especially so when the information needed is not available from other sources. Perhaps an example will help.

Surface mining procedures have become very widespread in the last few years with the emphasis upon the use of coal as a major energy source. During this same time, federal and state controls on surface drainage systems and water quality have become more strict. A problem has arisen in that much of the existing knowledge for construction of drainage systems and sedimentation basins for surface mining operations is based on theoretical models and computational algorithms developed for agricultural activities on generally flat topography. Consequently, there are many serious methodological problems involved in extrapolation of these agricultural methods and models to mining operations in areas of great topological relief. The appropriate adjustments and modifications of models and computational algorithms, initially designed for flat land agricultural drainage and sedimentation problems, did not exist. Consequently, it was difficult for mining engineers working in the coal industry to design proper temporary stream channels and storage basins to insure compliance with Federal and state water standards concerned with stream loads and erosion-deposition standards.

An early response to this need was research by agricultural engineers concerned with the proper modifications and elaborations of the earlier agricultural models to make them more appropriate to surface mining applications in high relief topography situations. This required extensive theoretical and empirical modeling, the collection of much existing data on rainfall characteristics, soil properties, topography, and other major variables. All of the information from these sources had to be integrated and presented as a series of computational algorithms, nomographs, charts, and procedural rules by which fruitfully and accurately to apply earlier models, such as the Universal Soil Loss Equation, to problem situations very different than those for which the models were developed originally.

This work took many years of effort by a few university researchers. As the research developed, it became apparent the new adjustments and modifications would be very helpful to practicing mining engineering operations. Consequently, graduate courses at a university began to be taught in this area. The professors involved refined their models and procedures and gradually produced a set of problems and notes which became a course on "Hydrology and Sedimentology of Surface Mined Lands." The course and the notes evolved into a textbook and an extremely popular continuing education short course for engineers concerned with constructing drainage system and sedimentation basins for suface mining operations (Haan & Barfield, 1978). The course is in demand because it represents knowledge and skill which is not otherwise easily attainable, but which nonetheless is central to proper complicance with sound practice and with state and Federal laws. The persons seeking this knowledge and skill include state and Federal inspectors as well as many individuals from engineering firms who design the temporary drainage and storage systems.

This pattern is not unusual in the development of continuing education courses. Such courses arise out of needs of practicing engineers for specific types of knowledge and skill. The courses developed to meet these needs are more likely to become effective if care is taken in assessing the effectiveness of the individual course in meeting these needs as the course is developed and refined through its various stages.

There are a variety of procedures which provide the information needed to improve developing courses through formative evaluation procedures. These include small trial offerings or pilot studies. These early experiences can provide much information about necessary revisions in course content, pace, duration, and presentation. Early informal contact with engineers, who are faced with problems in their work settings and who appeal to university or college faculty for assistance in specific areas of technical expertise, often precedes the development of a particular course or courses. In this way, needs of practicing engineers are often identified and later these needs may be met, in part, through a formal short course or another continuing education activity.

Selecting Course Content

One of the areas in which a pilot or trial use of a continuing education course can provide information is the appropriateness of the content included in the course. Content selection is always a problem. There is always more content than can be included in any course. Moreover, continuing education courses are typically of short duration, often consisting of intensive two or three day sessions, workshops, or sometimes weekly sessions for an hour or two over a period of several weeks. University professors often have a tendency to include large amounts of content in such courses. Participants often seek knowledge of much more limited information and procedures. Just how much, and in what depth, and in what breadth, should be included in the content of a short course is often not possible to determine until the course has been taught a few times. The impressions of the participants about how useful the content is in their work setting, as well as about the scope and pace of the course, are important. This information should be routinely sought in the early stages of course development. Questionnaires and interviews of participants before and after the course are useful to determine what they think they need to know and how much they think they have learned. Follow-up interviews and questionnaires to participants and their employers after completion of short courses are also important. Appendix A contains several actual questionnaires used for this purpose by the Learning Outcomes Measurement Project. These instruments may serve as useful examples to persons with an interest in formative evaluation of similar courses.

Oftentimes participants are introduced to methods and procedures in short courses which can become very useful to these persons when they return to their work setting, but only after much additional practice following the short course. If the procedures and methods are seen to be useful by the participants, and if these persons are convinced from course activities that they have the basic competence to apply and properly use these procedures, it is likely that the procedures will be practical and perfected upon returning to the work setting. This is particularly so if the short course is designed to provide the participants with a packet of materials and procedures to take back to the work setting. These materials and procedures can include computer programs, technical manuals, charts, tables, computational algorithms, and many more types of procedures or information which make the solution of certain problems easier and more technically correct.

When the outcomes of a short course depend upon such continuing use of skills and procedures, one should not expect participants to be completely facile with the skills and procedures at the end of the 3 day short course. In such a situation, assessments of the participants' knowledge of how to study further and independently use the procedures in the work setting and his or her willingness to do so are important factors. Subsequent information about the degree to which participants actually use the procedures and ideas taught in the course in their work settings, as well as information about the accuracy of the applications, is also very important. What one learns from evaluations of short courses by questionnaires and interviews with participants immediately following the course, and with participants and their employers some weeks after the course, is very important information. It can be used to change course content, teaching procedures, pace of instruction, and content organization toward developing a more effective course.

Refining Course Objectives and Learning Assessment Tasks

The objectives for courses also should be subjected to formative evaluation procedures. Sometimes objectives can be expressed in example problems or tasks which define clearly what it is the participant will be able to do when the course is completed. This is the case in the Haan and Barfield (1978) course on hydrology and sedimentology. The example problems and the problems to be worked at the end of each unit of the course are very clear statements of the types of skills and knowledge needed to solve this class of problems. Furthermore, the complexity of the problems cumulates until the problems in the final unit incorporate knowledge and skills from each of the prior units. This approach is very sound for it not only provides the participant with a concise and clear objective (the problem itself stated in its unsolved form) but it is also the means of assessing the competence of the student. The student's performance on these tasks is useful for diagnosing what has or has not been learned, and what needs to be done next in instruction. This approach is widely advocated for teaching complex technical content (Gagne, 1967; Gagne & Briggs, 1974; Webb, 1970).

The same problems, which functionally define the objectives for the short course by presenting the participant with problem tasks that he or she should be able to solve after completion of the course, also define a series of assessment tasks which may be used as pre-tests prior to the course to determine where learners are entering in terms of prior knowledge and skill. These same tasks can also be used as post tests after the course to determine exit levels of participants' knowledge and skill. They may also be used as embedded assessment tasks during the actual course of instruction to diagnose any particular learner's need for additional instruction on any particular point. The same sample problems also may be used in follow up studies long after course completion, to determine the degree to which course content and methods are being applied appropriately in the work setting. To the degree that the sample problems in the course are representative of the real problems faced by the practicing engineer, actual samples of persons' work in their job settings may be examined and scored on the incorporation and correct use of course concepts and methods. This type of assessment of learning outcomes for a continuing education course is the most rigorous possible. It provides information, not only on the degree of student learning, but also on the degree of course effectiveness for different types of participants with varying levels of prior experience and education.

Traditional behavioral objectives are generally not particularly useful for these important instructional organization and learning assessment functions, unless they are stated in terms of classes of performance outcomes rather than as highly specific entities or "behaviors". The outcomes of short courses, although narrow and focused, are not usually specifics, but usually an area of general skills or a class of performances. For example, typical outcomes include the proper design of experiments, the construction of runoff control and storage systems, or the assembly of information and decision making devices from micro-processing electronic components and the interfacing of these with industrial machine systems. In each case it is not a particular set of specific behaviors which are the intended outcomes of instruction, but rather a class of generalizable performances. No two experiments are identical, nor are any two drainage and runoff systems, or any two industrial production control systems. The desired outcome in each case is a set of basic but generalizable skills. Having learned such a set of generalizable skills, the course participants should be able to better design any experiment in a wide range of situations; design better runoff drainage and storage systems across a wide variety of soil, topography, and climatic conditions; and use micro-processing equipment to control a wide range of machinery used in many different industrial production operations.

The best route to this outcome is to insure an adequate sample of different types of problems in which to train participants to a criterion of mastery in the performance, with special attention to inclusion of a sufficient range of problem conditions so as to alert the participants to the typical adjustments which must be made in theory, assumptions, algorithms, or methodology. There is much support for this approach to instruction (Bloom, 1976; Carroll, 1963; Gagne, 1967; Gagne & Briggs, 1974; Manning, 1970).

There are always some minimum number of experiences which the participant needs to encounter across some range of variable problem situations if the skills are to be learned in this generalizable fashion.

Formative evaluation of programs and courses can help determine this optimal array of problem situations. It should also be noted that the tasks used for testing knowledge and skill of participants should be similar to the tasks and problems from the actual domain under study. These test tasks ought to be no different from the tasks used for instruction, except that they have been reserved for testing and they contain new problem situations not previously encountered in this specific configuration. It is important that the test tasks are not the same problems as presented in practice and demonstrations in the course of instruction. The real problems the engineer faces in his or her work setting will be similar to but not identical to the problems selected as instructional activities. By having similar, but different test problem tasks, the course participant's ability to abstract and generalize the general principles and procedures presented in course learning activities to other problem situations, not before encountered, can be assessed. This produces a better estimate than would otherwise be obtained about how well the engineer is able to transfer material learned in the course to real world problem situations.

Test tasks frequently need to be abbreviated, with part of the problem being worked out, or the problem being described and alternative approaches being presented, the course participant having to choose the most appropriate approach under the conditions stated. Such items can test for high levels of comprehension and skill but not require as much time as would working out an entire real problem. More will be said about the construction of this type of test tasks in latter chapters. (An example of this type of abbreviated test tasks designed for an actual continuing education course in hydrology and sedimentology of surface mined lands is found in Appendix B). Of course, what is lost with an abbreviated test task is a certain degree of validity and thoroughness in the assessment. There is no substitute for assessment based on actual evaluation of real work performance but there are good approximations which can be done more easily and more efficiently, and which will indicate the presence or absence of minimal comprehension, skill, and ability to solve typical problems encountered in the work setting.

Measuring Only That Which is Appropriate

It is important to note that one does not need to continue to collect all of the information one can about a course after that course is reasonably well developed and shown by its formative evaluation to be shaped in the direction in which it operates most beneficially for participants. For example, in the early stages of developing a course, questions concerned with the appropriate organization and pacing of content, the utility of the course material for practitioners, and the correct emphasis and amount of content are all of primary interest. After several replications of a course and the collection of this information through the routine procedures described earlier, the course may be thoroughly developed and function well. If this is the case, there remains only a need to assess the learning outcomes of individual participants on a regular basis for purposes of reporting to them and their employers the progress of individuals. There is no need to do the extensive assessment of how appropriate the course objectives and content are, or how effectively the course is operated and presented. Rather, over many replications of the course, these other questions can be asked and answered through appropriate observation and data collection procedures on an occasional basis.

One of the most effective means to monitor the quality and effectiveness of a well developed course over many replications is the method of "item sampling" (Lord & Novick, 1968; Shoemaker, 1973). Under this method the many questions and test items which are useful to evaluating course effectiveness are broken up into small sets. Course particpants are randomly selected to respond to one small set of questions or test tasks. Over replications of a course, much information may be gathered about all aspects of the course with minimal demands upon the participants time and energy.

The same principles apply to instructors of short courses. Once a given instructor or instructional team has successfully demonstrated that the course is taught in an effective manner, instructor performance need not be monitored very carefully during each subsequent replication. Some form of routine and brief participant evaluation of the instructor(s) should be continued to insure that participants have the opportunity to communicate suggestions and criticisms to instructors and to maintain instructor awareness and sensitivity to student needs. If a wide variety of information about the instructors is desired, a very long instructor evaluation form can be broken up into 3 to 5 shorter evaluation forms. These multiple forms would each contain different items. Course participants would be randomly assigned one of the forms. Over replications of the course a great deal of information about participants' judgment of the instructor(s) can be gathered, again, with minimal expenditure of time and energy. In fact, item sampling procedures usually yield more information than traditional procedures where all students complete all items. This is because the number of items included in item sampling assessment measures can be much greater in total than is possible in traditional assessment procedures. This means that a broader range of the domain of interest may be assessed.

Of course, item sampling does not work well unless there are many students involved. Many replications of the same course provide adequate numbers of students. There is a second limitation. Item sampling procedures are very useful for making judgments about course effectiveness and the skill of instructors. However, since not all students are tested upon or asked to complete the same questions, inferences about individual student's learning achievement and attitudes toward the course are not possible. For this reason item sampling procedures are useful for evaluation and monitoring courses and their effectiveness. They are not useful for making assessments of individual learners' achievement. Therefore, it is usually wise to retain some form of common but abbreviated test tasks which are administered to all participants, usually in the form of a test. Later chapters deal with the procedures appropriate to developing and using tests and other assessment procedures for the purpose of measuring and estimating the degree of individual student achievement.

These same principles apply to the collection of information about course participants. Initially it is of great importance to know much about the participants who are likely to be involved in a particular short course in the future. The best way to do this is to monitor present enrollments on the variables of interest. Important variables concern the diversity of participants' expectations, prior levels of skill and knowledge in the prerequisite skills and content for the course, and participants' occupational history and present level and area of professional activity in engineering. If the participants in a given course are extremely diverse with respect to necessary levels of prerequisite skill and knowledge, the course will be very difficult or even impossible to teach in an effective manner. If a course is geared at too low a level or too high, if it is paced too fast or too slow for a large number of participants, it will encounter difficulty. Thus, early attempts to learn much about the specific characteristics of the population of enrollees for a course are important.

Once the characteristics of the population which is to be served are known, adjustments can be made in the content and pace of course offerings. Sometimes it is necessary to meet diverse needs of participants by planning and offering more than one course, some at a very technical and advanced level and others at a more basic level. Again, data collected routinely in the early replication of a continuing education course, which initially provided information needed in the formative evaluation of the course, may not need to be routinely collected after the course is well established and has a stable population of enrollees with similar characteristics. The goal should always be to collect only that information which is necessary to make adjustments which are needed. Testing and other assessment procedures, such as interviews, questionnaires, and evaluation of on-the-job work performance, are costly and time consuming. After a course has been well developed it is reasonable to sample replications of courses and participants within courses on specific questions to achieve good information about the ongoing quality of the course and its outcomes. Time is always limited and most of the time available for short courses needs to be devoted to instruction.

This is not to say that testing or assessment of participants' knowledge and skill in the area of course content should not continue as courses are developed. Continual use of pre-tests, test tasks embedded in the learning activities of the course itself, and post tests can be very important instructional methods. Proper use of these methods can be very helpful to participants and instructors in producing better learning outcomes. However, if such testing is to continue, its purposes ought to be directly related to instruction by letting learners and instructors know the entry level knowledge and skill of particular learners, the types of practice and assistance most needed by particular learners to master certain areas of the course, and to report to the learners and their employers the amount of learning which has occurred after the course on certain specified outcomes.

Packaging and Delivery Considerations

There are many ways to package and deliver short courses. Some of these include on site training by an experienced professional; the use of specially designed textbooks, workbooks, programmed learning manuals, and other printed materials; the use of demonstrations, films, audio and video presentations, often with ancillary printed materials and homework exercises; laboratory demonstrations and activities (such as the assembly of components designed to perform in a certain way after first receiving lecture and individual instruction in the basic principles of the components and the uses to which they may be put); and working of sample and demonstration problems in groups or individually with the assistance of instructors and other experts.

The packaging and delivery of a course depends on a number of factors including the nature of the material to be taught; the geographic distribution of the participants; the need for specialized equipment such as computers or laboratory equipment; the complexity of the material to be learned; the level of prior knowledge and skill required of the participant; the availability of quality instructors; and the expected duration or "life span" for the course.

Some things cannot be learned from books or manuals. Some particular types of courses need to be taught as supervised experiences much the way surgeons are taught the finer points of various procedures. Other information can be readily taught through lecture accompanied by appropriate charts, graphs, and illustrations and followed by practice exercises. Still other skills for other courses can best be taught by individual study of printed materials and the working of rather traditional format problems presented at the end of sections of reading.

If the potential participants for a particular short course are widely dispersed throughout different companies over a wide geographic region, it is likely that a short course conducted at a national or regional professional conference or a special regional conference of a few days duration will be most appropriate. In such cases, questions of the location and adequacy of facilities and the dates and times of the conference activities are important variables which need to be considered. Not all times, locations, and facilities may meet the needs of the participants. Conversely, if there is a large local population of engineers needing a particular area of training and all of these individuals are located primarily within a few nearby companies or consulting firms, local conferences or even extended programs meeting weekly for a period of a few weeks are viable options.

It is often more cost effective to package a course so that it can be mailed out and used locally by any of a number of organizations and groups with any of a large number of instructors, once a course is carefully developed and found to be effective. A common method is to package the program of study in the form of mail-out video cassette lectures and demonstrations coupled with the appropriate ancillary printed materials, workbook, and problem materials. The initial cost of preparing the course in this form may be high. However, the advantage is easy replicability at many sites. If a course has to depend upon one person or a small group of persons for its instruction, it will be very limited in effectiveness by virtue of the available time and energy of the instructor or instructors.

One good example of a course which is packaged in a very cost effective way insuring wide replicability is the "Design of Experiments". This course was developed by Dr. John Van Horn of the Westinghouse Corporation and is taught by Professor J. Stuart Hunter (Box, Hunter & Hunter, 1978).[3] This course is very popular, since it deals with functional skills of experimental design, a topic central to the work of many practicing engineers. The potential audience for the course includes engineers and other technical staff in many industrial firms scattered throughout the country. The packaging of the program in video cassette lectures and demonstrations coupled with printed materials for individual participant study is a highly replicable format. Groups of engineers at any location may convene, have the video cassette mailed out to a central person, engage in individual study of the materials and working of the problems provided, and jointly participate in watching and discussing the lectures and demonstrations on the video taped programs.

Before a course is this fully developed and packaged it should have been well evaluated and modified to be very effective as determined from this early formative evaluation activity. Moreover, this much effort and money should not be expended on a course unless it is a course which is likely to continue to be needed by large numbers of participants into the future. It is also important to recall a point made earlier. After a course such as the "Design of Experiments" is carefully developed, evaluated, and found to be effective, there is little need to continue the same intense level of evaluation activity. Rather, it is appropriate to sample some particular instances of the course and some participants to determine if the course continues to meet the needs of participants and to obtain information on possible revisions for future versions. Evaluation of individual participants' learning is still appropriate as a continuous aspect of the course in order to communicate to the learner and employer an estimate of the degree of initial learning resulting from participation in the course.

---

[3] The course is owned and distributed by the Office of Continuing Education and Extension, College of Engineering, University of Kentucky, Lexington, KY 40506.

Selection of Instructors

Part of the packaging consideration concerns selection of an instructor(s) for the teaching of the course. Certainly technical expertise and competence in the content area of the course are basic criteria. However, these are not sufficient. The most appropriate instructor may not necessarily be the most expert university professor, but some other skilled practitioner. It is not that university professors should be ruled out as appropriate instructors. A great reservoir of talent resides in such persons. Rather, it is crucial to select from among those professors the individuals willing to place the needs of the participants first rather than the objectives of the instructor. It is important to select those professors responsive to the needs of participants rather than those bent on achieving their own dearly held objectives regardless of participants needs or concerns.

In practice it often turns out that what practicing professionals need to perform better is something other than what their professional societies and leading academics think they need. It is also often the case that the practitioner needs a broadened understanding of theory and principles in order to perform his or her work more wisely and efficiently. In such situations it is the job of the short course instructor to incorporate this theory and increased understanding by careful selection of examples and illustrations which clearly demonstrate how a knowledge of the broader theory and relationships makes it easier to solve the problems of practice in a more sound and integrated way. Sometimes this can be done with anecdotes of stupid, dangerous, or uniformed practice on the part of a person who sees only a very little of the problem area for lack of a broad enough grounding in relevant theory and concepts.

These anecdotes can sometimes help participants understand the importance of theory to practice.

One example from another technical field is the medical laboratory technician who called the repairman for the flame photometer each week because the sodium readings on blood serum samples and known calibration samples were highly erratic. After a period of six weeks, during which the equipment had been condemned as faulty, the technician in preparing to wash glassware was observed energetically shaking large amounts of soap powder into water in a sink near the photometer. Soap dust was everywhere making everyone sneeze, settling on the funnel in the flame photometer, and being aspirated with the next sample, resulting in a very high sodium reading. The technician apparently thought sodium was something peculiar to human blood serum and fluids, and made no connection between the soap powder and the high readings. Yet the same technician was extremely careful to rinse all glassware three times with distilled water to insure no contamination of equipment. There must be many good examples of too narrow an understanding of engineering principles which lead to inappropriate practice. Instructors of courses for graduate engineers ought to be able to provide many similar examples and convincing arguments that oftentimes the best route to better practice is a more solid understanding of theory and major concepts.

Specialized Equipment and Facilities: Implications for Course Delivery and Learning Assessment

Still another consideration in the packaging of continuing education courses is the need for specialized equipment and facilities. For example, a short course on the latest developments in microprocessing equipment and the use of that equipment in the control of industrial machining processes may require special equipment and facilities. First, Heath kits or some similar self-instructional packet of equipment may be useful and even necessary as an instructional activity within a laboratory setting. Second, the availability of a central computer facility with the capability of simulation of industrial systems control processes may be necessary in order that participants may test and revise the logic of the programs they prepare for their control units. Third, an actual field trip or demonstration of a number of current applications of microprocessor control systems to industrial production process may be desirable.

On the other hand, if one is teaching a course on general principles of engineering economics, the only facilities and specialized materials needed may be a text book, a set of programmed instruction work sheets and problems, and some common lectures for participants. The lectures can be presented live, filmed, or on video tape. They can be disseminated by mail, telephone lines, or communication satellite, as was the case in a recent project in the Appalachian region (Mertens, 1979).

In short, the need for specialized equipment and personnel as well as the complexity of the material, dictate much about the packaging of the course, including such things as the mode of instruction, laboratory, lecture, workbook, fieldtrip, and combinations of these; appropriate types of instructors, university professors, practicing engineers, or other specialists; location of the course near central facilities or dispersal of the course throughout a wide geographic region by use of printed materials, films, instructional kits, and other media; and the number of replications anticipated for the course in the future.

The evaluation of learning outcomes for different types of course delivery and packaging are also different in a number of ways. If one is offering a course on the assembly and use of microprocessors it is foolish to test only with a paper and pencil test. Much more appropriate and useful in diagnosis of the learning outcomes is some form of practical laboratory test, usually shortened and abbreviated in scope, but nevertheless sufficient in assessing basic skills and knowledge in a performance area. Yet in a course on "law and engineering", a pencil and paper test consisting of multiple choice items which require the participant to make judgments about the legality of certain engineering procedures in industrial manufacturing in specific cases presented in individual items, is a perfectly valid approach and may provide the means to assess a much wider range of knowledge and skill than would a more performance oriented, practical examination.

Practitioners' Tacit Evaluation of Courses

The most convincing data on the degree of individual participant learning is improved performance in job activities. It matters very little whether or not the engineer's improved work performance resulting from the course is actually formally measured as long as employers and their employees who have completed the training observe that improved performance results. As mentioned earlier, it is often such informal tacit evaluation of the effects of continuing education courses which is the most functional type of evaluation. If past participants and their employers see evidence of the utility of a course in improving skill and performance of practicing engineers, the word gets around and the course becomes heavily subscribed in the future. If this utility is not perceived, no matter how "proper" the formal evaluation of the course and its effectiveness, the course will not be heavily subscribed.

The point is that in the development of such courses, the tacit evaluations of participants and employers concerning the utility of the course ought to be sought and attended to. One can argue that this is an attempt to "please" the client, a tactic which some academic professors of specialized engineering sciences do not approve. However, it must be remembered that the companies and individuals who enroll in continuing education courses in engineering do so only by virtue of taking time off from work, at corporate and personal cost in terms of dollars, time, and energy; and in the pursuit of meeting particular types of needs related to improved work performance. Therefore, what the individual participant and his employer want, and especially what they think to be the value of a particular continuing education course, are the major criteria for course effectiveness.

Conclusion

All of the previous material should make it clear that one does not simply evaluate the learning outcomes for participants for any continuing education course. One must also attend to and carefully evaluate the characteristics and effectiveness of the courses which are developed to meet the needs of practicing engineers. The functional learning outcomes of any course depend not only on the learner, but to a large degree, upon the structure and organization of the course. Careful formative evaluation activities, like those described in this chapter, help insure that a course will eventually be well developed and result in significant learning by individual participants. Without this early formative evaluation and subsequent course revision, it is doubtful that even elaborate attention to the development of "tests" of participant learning resulting from the course will be meaningful. When any population of students seeks out instruction to meet their own perceived needs in specific areas related to their work and at personal cost to themselves, the major criterion by which the instructional experiences will be evaluated is the perceived utility of the course to the work performance. The perceptions of participants and employers on this matter of perceived utility are paramount. Proper attention to formative evaluation procedures in the development and packaging of continuing education courses can make it very likely that particular courses will, indeed, result in improved practice in some job related area.

After all of this is accomplished, it is appropriate to attend to the formal and routine evaluation of the learning outcomes of individual participants who have enrolled in and completed particular short courses and other continuing education experiences.

The next section of this document deals with the development and use of summative evaluation procedures, the means by which the progress of individual learners are reported to them and their employers, and also the means by which the general effectiveness of particular courses and programs are reported to professional societies; governing, academic, and standards boards; and others having a strong interest in the quality of individual courses in the continuing engineering education arena.

Chapter 5

SUMMATIVE EVALUATION OF LEARNING OUTCOMES

Earlier portions of this book have argued that the learning assessment procedures should be tightly integrated into the development and ongoing operation of the course. This is because the best tasks for assessing learning outcomes of students belong to the same universe of tasks from which instructional activities and performance objectives are sampled. That is, even if it is clear in general what is to be taught in some complex performance area, one still always must sample some finite number of specific topics to be studied, learning activities to be performed, and sets of instructional materials and methodologies to be used. When one does this sampling, one trys to select an adequate range in topics, materials, and activities to insure that the learners can generalize the knowledge and skill they gain from the instructional experience. It is not possible to teach all instances and applications of any particular knowledge and skill which may be the intended outcomes. It is only possible to sample examples wisely, and to provide the student with an appropriate breadth and depth of how the knowledge and skill may be usefully applied. If this sampling is done well it is likely that, having completed a course of instruction, learners will transfer or generalize the knowledge and skills acquired, through exposure to a finite number of carefully selected learning experiences, to many other situations related to their work activity and job performance.

Because of the complexity of developing courses it is best to refine early instructional organizations given information about the effects of the courses on learning outcomes of course participants. This formative evaluation is an ongoing process essential in the early stages of course development, but it remains important throughout the lifetime of a particular course offering.

Need for Summative Evaluation

There is, however, a second major purpose for evaluation. There comes a point when a particular individual who has participated in a course wants to know how much and how well he or she has learned. Employers who send engineers to participate in continuing education courses also want to know, in specific terms, if their employees have learning anything and if so, how much. In addition, course instructors are frequently asked to certify that students have acquired a set of skills or performance capabilities. They too need information about the degree to which individuals have learned the material and skills taught in particular courses (Enell, 1980, pp. 186-187).

No matter how well the formative evaluation activities are carried out in the development of a course, and no matter how generally effective a course is shown to be in terms of its overall effectiveness (summative evaluation), there remains a need to know something about the degree of each individual's learning following completion of a course. Therefore, it is necessary to assess individual learners' achievements in any course in order that learners themselves, or others designated by the learners, can be provided with specific information about how well the course operated for individual persons. In short, persons who attend courses, their employers, and perhaps the professional societies to which the individuals belong, often want information about the amount of learning which has resulted for participants in a continuing education course.

The common way to obtain such information is through various forms of testing. Alternative performance assessment methods also exist. The information gathered from such procedures is never exact in terms of reporting the degree of learning of participants. As argued earlier such measures of learning outcomes are only estimates. However, if tests and other performance assessments procedures are properly designed and administered, they can provide good estimates of the degree of individual learning on specific areas of knowledge acquisition and skill development. Furthermore, information of this type, collected and pooled across all individuals enrolled in a course, can be used to make statements of the general effectiveness of the course in achieving its intended learning outcomes with specific groups of enrollees in specific areas of skill or knowledge. When used in this manner the evaluation is called "summative" because the description of the course effectiveness summarizes its overall impact on learning outcomes of participants. However, the information of a summative evaluation can also be used in a formative way. Aspects of the course may be modified in future replications, given data on present effectiveness. When used this way the process is called "formative" evaluation because the future characteristics of the course are shaped on the basis of data about present effectiveness.

Differences Between Summative and Formative Evaluation

Summative and formative evaluation procedures differ primarily in how and for what purposes the information obtained from assessment activities is used. Formative evaluation has as its central concern the adaptation and modification of early and ongoing instructional design and actual operation of a course toward improving course effectiveness. Summative evaluation has as its main focus the reporting of the success of individual course enrollees in achieving specific knowledge and skill at some point in time after the completion of a course. Summative evaluation is also concerned with reporting the typical level at which a course achieves its intended learning outcomes for its enrollees at some given point in time following instruction.

Four General Learning Assessment Procedures

The next four chapters' or Part II of this book, detail different types of testing procedures useful for assessing learning outcomes. The testing procedures presented are used for both formative and summative evaluation purposes. Therefore, the procedures are presented in relation to the assessment of the degree of individual student learning and also as a means to estimate the general effectiveness of a particular course in achieving its intended outcomes with a particular group of enrollees.

The four testing procedures presented include pre-tests, tests administered prior to instruction; embedded tests, tests included in the course of instructional activities; post tests, test administered following instruction; and delayed post tests and other performance assessments, administered long after instruction when the individual has had time to actually apply and use in his or her work activity the knowledge and skill encountered in the course. Other learning assessment procedures, in addition to testing, are also presented. Different types of tests and their multiple purposes and uses are also discussed in this major section of the text.

Developing Sound Learning Assessment Procedures

Part III of the text consists of three chapters concerned with alternative methods for developing effective testing procedures for measuring learning outcomes. Chapter 10 explains in detail how to develop and use tests and other performance assessment procedures in the process of formative evaluation of courses. The procedures presented offer a means to develop good courses and good learning assessment procedures through an integrated process of developing instructional objectives, content, and methods along with the learning assessment tasks. The main purpose is to design valid learning assessment procedures which will properly serve formative and summative evaluation procedures.

Chapter 11 describes procedures for conducting test item analysis and test reliability studies to insure high quality learning assessment instruments. Duplicate procedures are provided for two common approaches to the evaluation of learning outcomes from instruction. These are the norm referenced and the criterion referenced (or mastery learning) approaches. The latter approach is the one recommended for continuing education courses for engineers for a variety of logical reasons which are presented.

Chater 12 details limitations of tests, no matter how well they are designed. The purpose of this chapter, along with other sections in the remaining chapters, is to prevent the abuse or misuse of testing in the assessment of individual's learning and in the formative and summative evaluation of learning outcomes for courses and programs.

Chapter 13 is Part IV of the text. It details how to collect, integrate, and report data gathered from multiple measurements of learning outcomes for purposes of formative and summative evaluation of courses as well as for reporting the degree of learning of individual students.

Collectively these remaining chapters provide detailed procedures and examples for developing sound learning assessment procedures for whatever purpose, be it formative evaluation toward course improvement, summative evaluation to describe overall course operation and degree of effectiveness, or reporting of individual learner achievement to course participants and persons they designate.

Chapter 6

PRE-TESTS, THEIR PURPOSES AND USES

Pre-tests have three basic uses. First, they may be used to inform prospective enrollees of the content of the course and the necessary level of prerequisite knowledge and skill required for successful completion of course activities. Second, they can provide course developers with information about the general entry level knowledge and skills of enrollees, information useful in making subsequent adjustments in course content, rate of presentation, and emphasis. A third function is to provide baseline information for each learner from which to make inferences about the amount of learning of individuals and groups through subsequent comparison with the results of performance measures and tests administered, during or following the course. Each of these three functions of pre-tests will now be examined.

Pre-tests as Informative and Screening Devices

Pre-tests can be used in a screening fashion to insure that persons who enroll for a course meet the necessary prerequisite level of knowledge and skill. For particular courses, short pre-tests can be mailed out along with course announcements to allow prospective enrollees to make their own private assessment of their readiness for a particular course and to decide what to do to prepare for a course to insure readiness for its learning activities. Other methods of informing learners of the prerequisite levels of knowledge and skill also can be used. These include a simple statement of what types and levels of knowledge are required, or a listing of particular prerequisite courses or experiences which should have been completed. The listing of prerequisites is a common practice in engineering continuing education.

Whether or not to use a pre-test as a screening device depends upon the nature and content of the course. In Chapter 2 continuing education courses for engineers were divided into four categories. These included courses with a focus on a) upgrading and remediation of basic knowledge and skills; b) broadening and extending previously learned scientific and technical concepts and skills; c) imparting new and advanced levels and skills in specialized technical areas at very high levels of expertise to keep pace with new developments in technology; and d) introduction to new areas of knowledge and skill outside of basic engineering practice such as economics, management, human relations, community development, and environmental protection.

A simple brochure describing the purpose of the course and the nature of the content and learning activities is probally sufficient for courses in the "d" category. This may also be true for courses in category "b" as well. However, in category "c", where there are expectations for very high levels of technical expertise in given areas prior to course entry, it may be reasonable to design and mail out a short pre-test which actually tests for this prerequisite skill and knowledge. A scoring key can be provided the prospective participant. After completion of the pre-test, he or she can decide if the course is too elementary or too advanced. Suggestions can be included in the materials sent to prospective enrollees concerning how to meet prerequisites. These would often include studying certain procedures published in specific journal articles or manuals; working with particular types of problems or equipment; and completing other specific short courses or formal learning experiences to master required basic concepts and skills.

It is also reasonable to use a pre-test in relation to courses concerned with upgrading and remediation of basic skills and knowledge. In this case, a pre-test can either be mailed out to prospective enrollees or administered at a common location prior to the course. If security of test items is an important consideration, the latter alternative should be used. Here the pre-test should consist of a set of items similar to those included on the "final examination" for the remediation course. In situations where the remediation course is designed to prepare students to pass licensing examinations, both the pre-test and the course final examination should be a sample of items typical of those found on the licensing examination itself. It may also be advisable to include a sample of diagnostic items on the pre-test to assess each individual's knowledge and skills in basic mathematics and physical science concepts which are required to successfully complete the problems presented in test items on the licensing examination. If the pre-test items are properly sampled from across basic diagnostic areas and the content of the remediation course, the prospective participant, as well as the instructor, can determine whether he or she needs to invest the time and effort in enrolling in the course. If the person performs very well on the pre-test, enrollment in the course would probably be non-productive. On the other hand, if the individual performs poorly, he or she would probably be well advised to enroll in the course. These procedures can be particularly helpful to persons preparing for professional licensing examinations.

Pre-Tests as Devices for Adjusting Course Content and Operation

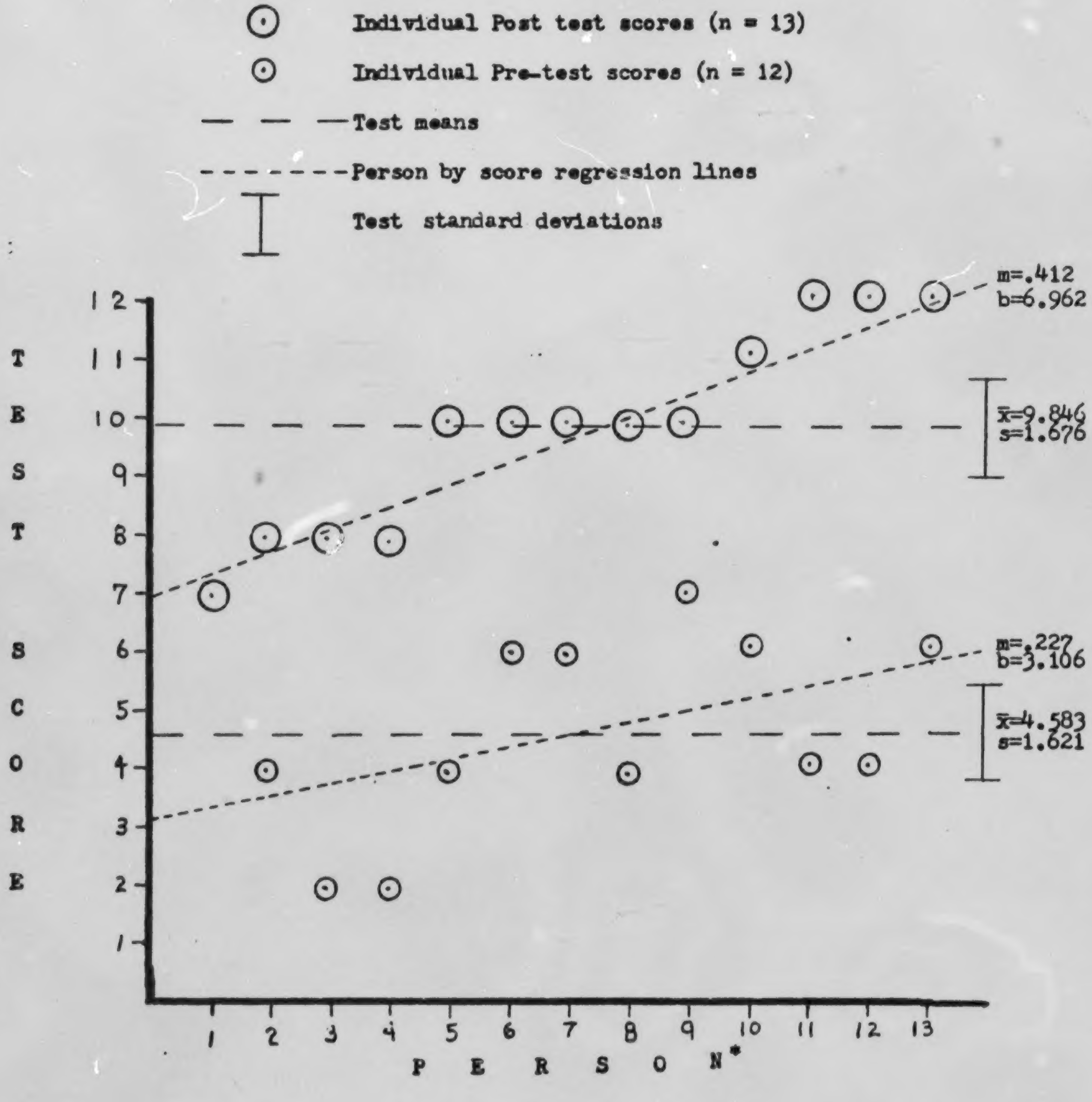
Pre-test information collected routinely on course participants at the beginning of the course can be very useful in providing information on any adjustments which need to be made in the focus of instruction, the rate or pace of presentation, and the level of complexity and difficulty of course content. Much has been said about these formative evaluation functions earlier. Additional examples of how to use pre-tests for this purpose are presented in Chapters 11 and 13. Here it will suffice to add that the information collected from the administration of short pre-tests at the beginning of courses across replications is very useful for this purpose. Pre-test and post test scores for a course may be measured and plotted similar to the example presented in Figure 1 in this chapter or Figures 2, 3, and 4 in Chapter 13. If similar plots are made over several replications of a course, and if the tests are valid and reliable measures of specific learning outcomes, much information about a course and its effectiveness under different conditions and with various instructional adjustments in methods, instructors, duration, and other factors can be determined.

Pre-tests as Indicators of Baseline Performance

The third function of pre-tests, to provide baseline information about the amount of knowledge and skill with which participants entered the course, makes it possible to compare the course participants' performance on post tests with the earlier measures. An example may help illustrate this point. Figure 1 is an actual plot of the

Figure 1

Pre- and Post Test Scores by Persons With Test Means, Standard Deviations, and Persons by Score Regression Lines for a Short Course



*Persons ordered by rank of post test score pre-test and post test scores of 13 medical laboratory technologists enrolled in a short course. The course was a six hour intensive workshop activity presented in two major sections over the course of one day. The course content was principles and techniques for enhancing student motivation and achievement in technical courses through use of appropriate instructional designs and teaching methodologies. All 13 persons were instructors of technical and clinical courses in medical laboratory technology in college and university programs.

The pre-test and post test score of each person is plotted against the rank of that same person on the post test score. A quick glance at the graph shows each participant's entry level knowledge of course content as measured by the pre-test. Also shown is each individual's exit level knowledge as measured by the post test. The horizontal lines represent the means of the pre- and post test scores for the group. These mean scores clearly illustrate that the performance of participants in the course improved following instruction. In addition, a regression line may be fitted to each set of scores. The slopes of the two lines reveal information about the differential effectiveness of the course for persons of differing ability levels as defined by their rank on the post test or by some other functional performance criterion. Statistical significance tests may be performed on the differences between pre- and post test means for repeated measure situations. In addition, statistical significance tests may be performed on the difference in slope of the regression lines for the pre- and post test scores. These statistical inference procedures can lead to powerful generalizations about the general effectiveness of any given course in improving the performance of participants. Replication of results such as those shown in Figure 1 across many trials of a given course are even more convincing of the effectiveness of a course than are the many statistical inference procedures which are possible. Chapter 13, pages 270 through 280, contain the pre- and post test results for groups of engineers enrolled in a short course on urban water quality modeling. Figures 2, 3, and 4 provide data for three replications of the course. Inspection of the graphs shows that the course is effective and consistent.

If no pre-test had been given to the participants shown in Figure 1 or to those shown in Figures 2, 3, and 4, it would have been impossible, on the basis of the post test alone, to determine the amount of learning taking place or even if learning had occurred. The soore on the post test would be uninterpretable because one would be unaware of the entry level ability of participants. In the absence of the pre-test, the inference could be made that the post test was too easy and did not measure the learning which had occurred. Another inference could be that the post test was of appropriate difficulty and that persons had learned from the course. Still a third inference could be that the post test was of appropriate difficulty, but that the participants were already knowledgable of the content of the course and might have scored this high or higher on a pre-test. This last inference would cast doubt on the appropriateness of the course for the participants. The point to be made is simply this: without the pre-test data it is difficult to choose from among the alternative interpretations of the post test results. Of course, a fourth inference could be that the post test score is unrelated to what the course participants learned. The validity of the test is questioned in this case. It goes without saying that any inference about how much persons have learned based on test scores needs to be derived from valid and reliable tests. Chapter 10 details methods by which to insure that tests are developed which are reasonably valid and reliable. Chapter 11 points out the limitations of even the best tests and provides suggestions for broadening one's inferences about learning outcomes by using other appropriate and multiple indicators.

In reality, there are always other indicators of the degree of participants' entry knowledge and the amount of learning experienced during a course other than pre- and post test score comparisons. These include: the amount of learning course participants report, the logical determination of the appropriateness of the post test or other learning assessment used by the course instructor, the rate and accuracy with which participants are able to perform the instructional activities and exercises during the course of instruction, and other similar indicators such as the ability and willingness of the course participants to use properly the knowledge and skills acquired in their daily work activities. Both earlier and later chapters of this book make it clear that these other indicators of learning should be systematically collected and used in making judgments about the effectiveness of a particular course as well as about the degree of learning achieved by individuals. However, this does not negate the need for more formal assessments of entry and exit level knowledge and skill. Information of the type presented in Figure 1 is very helpful, not only to the persons who operate the course and are concerned about its general effectiveness, but to the individual learner who can determine his or her own amount of learning as measured by the pre- and post tests, in relation to other course participants' performances and in relation to some level of mastery of the course content and skill as defined by the course developers or some other criterion such as common standards of practice (Lacefield, 1980).

Learning Resulting from Pre-test Experience

Generally if pre-tests are to be used, they ought to consist of individual items or assessment tasks parallel to but different from the items or tasks used in the post test. One does not care if students learn from experience with a pre-test. In fact, this is often a beneficial outcome. If the pre-test helps participants better understand what it is they need to know or what it is they need to learn to do, it can be very facilitative to the instructional activity which follows. Pre-tests can often define for the learner, in very operational terms, what needs to be attended to most in the instructional activities which follow. However, different items and assessment tasks are required for the post test because one does not want to be measuring, only or largely, increased knowledge and facility with specific test items or specific assessment tasks.

If one administers a pre-test, corrects the test, and reports the results to the learner, post test scores on the same test items given at some later time will be higher. This is generally true even if no instruction intervenes. What has happened in such cases is that individuals have learned specific responses to specific test items or performance tasks. They may have learned some other more generalizable things as well, such as how certain terminology is used, the style of the test developer, or obtained a better idea of what it is that the course instructor deems important enough to test for. Therefore, if one gives the identical test as a pre-test and post test, any gains in learning which appear are confounded. The confounding is between increased performance due to familarity with specific aspects of test tasks and items and increased general knowledge and skill in using and applying the concepts and principles learned in the course. Because of the confounding one does not know if the post test score represents only specific learning of how to get these items correct or generalization of basic knowledge and skill concerning ways to solve these types of problems. While the latter outcome is a proper goal of instruction, the former is not. Consequently, care must be exercised to make pre- and post tests parallel to one another, but to consist of different items. Otherwise the observed differences in pre- and post test scores will not be easily interpretable.

There are a number of evaluation designs which can be used to measure learning outcomes without recourse to pre-tests. Some designs even allow for the effect of learning from the pre-test to be separated from learning effects resulting from instruction as both are reflected in post test scores (Mason & Bramble, 1978). These evaluation designs will not be described here. However, they are especially useful when a course is to be replicated several times and different evaluation designs can be used with each replication to estimate various contributions to effects measured by post tests or other performance outcome measures. It suffices to conclude that, while it is not always necessary to use pre-tests, if properly constructed, pre-tests serve a number of important functions which make them worthwhile.

Two Approaches to Constructing Pre-tests

There are two general approaches to the construction of different but parallel forms of pre- and post tests. The first method is to define a sample of performances which the individual should be able to carry out after instruction. From this sample of performances, specific assessment tasks and test items can be prepared, without regard as to assignment to pre-test or post test use. In fact, as mentioned earlier, the listing of this sample of performances can also be useful for defining an appropriate range of learning activities and experiences by which to instruct the learner. In this sense, both test items, or other forms of assessment tasks, as well as the activities and exercises selected for inclusion in course content by which to instruct learners, are all samples of the class of performance capabilities. Learners should be able to apply the specific concepts, skills, and knowledge they have learned in a course to the successful completion of real life problems in the content area under study or to test items which simulate these real life problems.

Under such a plan some of these performance tasks are assigned to a pre-test role, others as post test, and still others as instructional task or practice activities. In such a case, the course designer will select several sets of tasks much alike in terms of the levels of difficulty, the types of concepts and skills which are being applied, and the types of problem situations they represent. The pretest, the sample of actual instructional examples and practice activities selected as course content, and the post test ought to all reflect a common breadth of difficulty levels and areas of application. In such situations, pre- and post test comparisons of participants' performances, as well as the accuracy and ease with which instructional practice activities are completed during the course of instruction, are very useful in making inferences about the amount of learning resulting from instruction. If such a plan is followed, it is important to insure that both the pre- and post tests contain a proper distribution of test items or assessment tasks. The problems presented in test items must require application of course concepts and principles across a range of typical conditions representative of those encountered in actual work areas.

In this first approach, there is no attempt to prepare duplicate pairs of items to serve pre- and post test functions. Rather the entire domain of performance being taught in the course in conceptualized as a universe of multiple performance capabilities. The course designer samples from the domain many specific tasks which collectively represent what it is a person must be able to do to exhibit mastery of the performance domain. Many individual items are developed, with attention to achieving a proper distribution of items across various levels of difficulty or complexity.

Assignment of items to the pre- or post test role is random with no particular attention paid to matching each item on the pre-test with an equivalent item on the post test. Taken as a whole the two test forms are considered equivalent because their items are drawn from the same universe, not because they consist of individual paired items. Items on both forms of the test are sampled in a uniform manner across the domain. Either form of the test may be used interchangably in any test role, pre-test, embedded test, post test, or delayed post test. This method works best when the domain being tested is very broad and the tests developed are quite long. This approach to parallel form test construction is very useful in courses in the first category concerned with remediation of basic knowledge and skill. Here the performance domain is very large and many parallel form tests can be sampled from that domain with little or no duplication of individual items.

A second general approach to the construction of pre- and post tests is to first define the content and skills expected as outcomes from instruction and to then prepare duplicate items for each area of performance to be tested. The items are duplicate in terms of difficulty level, knowledge required, and the specific skill or concept to be applied. They are different only in the specifics of a given problem or problem application situation. What results from such an approach is a parallel set of items; one set may be used for the pre-test and the other for the post test. If the pairs of test items are developed, the members of each pair may be randomly assigned to either the pre- or the post test. In addition, the entire item assembly in the form of the pre- or the post test, may be used for either function. That is, the two parallel forms may be used interchangably as pre- or post tests. They can also be used in other test roles such as embedded tests or delayed post tests. The basic plan under the second approach is to develop one good test which is comprehensive in terms of its representation of course knowledge and skills and in terms of a mixture of relatively easy though difficult tasks. When one selects a particular area of knowledge or skill outcome to be tested, one simply develops two items each time, rather than only one. One item is then assigned to the pre-test item pool and the other to the post test item pool. This second method is most useful when the domain is well defined, somewhat narrow, and the tests to be developed will be short, perhaps under 30 items or thereabouts. The second method is probably most suitable to the typical short course with its few well defined intended learning outcomes.

Either of the approaches described above will produce pre-tests which make initial assessments of participants' entry learning levels and post tests by which to make inferences about the amount of learning in specific areas following instruction. Such information is of value to the persons who operate the course, the course participants, and the persons who employ and professionally certify engineers.

Chapter 10 contains detailed procedures for the construction of pre-, post and other types of tests. Detailed procedures for the proper sampling of test items within the performance domain of interest are provided. Both methods of preparing multiple forms of tests are presented. Procedures for insuring the validity of the tests are described. Chapter 13 presents detailed information about how to use test data to make inferences about the degree of learning which has been achieved by individuals and to make judgments about the overall effectiveness of courses.

Chapter 7

EMBEDDED TESTS -- THEIR PURPOSES AND USES

Embedded tests are simply assessment tasks built into the instructional sequence and interspersed with other instructional activities. Common forms of embedded tests include the homework problems frequently assigned after each class session in typical courses in engineering and related scientific and technical fields. Frequently these completed homework problems are collected, corrected, the student's performance recorded, and then returned to students. Students may "cheat" on such homework assignments, but are foolish to do so because doing the problems is a practice activity, often the major learning activity, by which students become skilled in the application of course concepts and principles.

Traditional Embedded Test Tasks

Although homework is given for practice, the results of homework performance by students can be examined to reveal the degree of student understanding of course principles and concepts as well as the presence or absence of more basic prerequisite skills. Good instructors interact with students in precisely this manner. They frequently assign homework in graded levels from simple to more complex problems. They collect the homework, requiring that it be completed in a reasonable length of time. They then correct the homework, not only indicating if the problems are right or wrong, but pointing out errors to students. Frequently, the instructor identifies common problems which reoccur in a given student's work. The instructor often calls the attention of the student to the problem area or areas in written comments or by a personal conference. Sometimes it becomes apparent to the instructor that the majority of students have misunderstood or not fully understood some particular procedure or concept. Then the instructor often takes steps to correct or remediate the area of misunderstanding in subsequent lectures, laboratories, or similar class learning activities. Homework, therefore, is a form of assessing learning outcomes particularly useful in formative evaluation activities.

Although good instructors have been using these procedures for many years, it is important to realize that the data collected from the performance of students on such "embedded test" tasks can be used to make strong inferences about the degree of learning achieved by individual students and the general effectiveness of particular courses in achieving their intended outcomes. Other traditional embedded assessment tasks include frequent quizzes, practical or laboratory examinations and tasks which require the demonstration of a particular skill at various levels of safe and competent practice. The design and completion of various experiments, special projects, reports, and analyses are other examples of common embedded tasks. Each of these tasks calls for a skilled performance of some type. Often parts of the performance can be observed directly, particularly in the most critical stages. In addition, the performance almost always results in a product of some type, whether it be a set of design specifications for a bridge or a dam, a design for a scientific experiment, or an analytical report of the structural properties of a particular type of material which is to be used in the fabrication of machine parts.

The actual observation of the student's performance is a very informative assessment of the person's level of competence in using the skills and knowledge which have been taught. Examples of areas where such observational assessments are appropriate include: performances using special equipment such as scintillation counters, electron microscopes, or x-ray crystallography equipment. Other examples involve the selection and use of computational algorithms in solving particular problems; the making of certain assumptions about the way a problem task is fruitfully approached, including identifying those solutions which cannot be used; and the development and use of computer programs for data analysis and/or simulations.

The direct observational assessment of the student's performance (and the products resulting from the performance) to determine the degree of learning, and specific strengths and weaknesses in an individual's competence, is a powerful assessment procedure. This differential assessment of students' strengths and weaknesses in complex areas of performance based upon direct observation by experts who are also tutors is a form of embedded assessment task which ought to be used more frequently. These types of learning assessment procedures are very useful to instructors in diagnosing what it is in particular students are and are not yet able to do. In addition the assessment procedures are also very instructive to the students. Embedded test tasks comprise an integral part of the instructional activities by which complex performances are learned. Persons seldom master complex performances in one trial. Rather, many trials are required and performance may be expected to be very variable and incomplete across parts of the task, especially in early stages of learning (Bugelskik, 1971; Gagne, 1977). Embedded tasks of the type which have been previously described are very useful in testing the individual's performance capability across the various parts or component skills and knowledge required for skillful and accurate whole task completion. They are extremely useful for making decisions concerning when instruction can be stopped in certain areas of performance because the person has learned the skill or concept to a mastery level and further instruction would be pointless. Embedded tasks are also useful for determination of where additional instructio and practice are necessary.

Embedded test tasks may take many forms. As we have seen, they are often practice problems assigned as homework. They can be quizzes, projects, experiments, reports, and similar activities. However, embedded tests can also be more abbreviated samples of work performance as is the case with pre-tests, post tests, and delayed post tests. Although actual samples of performance under conditions similar to those in the work setting are the best ways to assess the degree of learning in a given area, other considerations often prevent the inclusion of very many such assessment tasks. Sometimes there is simply not enough time to provide the necessary number of indepth instructional and laboratory learning activities and also provide equally complex and involved assessment tasks by which to make inferences about the degree of student learning. One solution to this problem is to keep better track of each learner's success and lack of success with each practice and laboratory task which is assigned as part of the instructional sequence. Another solution is to use an abbreviated sequence of tasks as test items by which to make reasonable inferences about the degree of learning achieved by students on the complete performance task. The first solution has already been discussed in terms of keeping good records of student performance in the completion of instructional activities such as the doing of homework problems or the carrying out of laboratory activities. Let us now consider the second alternative.

Abbreviated Forms of Embedded Test Tasks

Even on regular tests in areas of complex performance, the test tasks must be abbreviated. Often students are not required actually to complete the solution of a complex problem, but only to set up the problem in a correct manner. This tests for the student's understanding of how to formulate the problem, what class of solutions to consider and select, as well as what mathematical models and procedures to use to achieve a solution. It does not test the student's ability to accurately complete the solution. Oftentimes, when it is of interest to test the students facility with computational procedures as well as problem formulation, the problems presented as test items are modified to be more simple than those usually encountered in practice. Values given are frequently in small or whole numbers. Actual computation is deliberately simplified to insure that the student can complete the problem in a short period of time, usually minutes, without the aid of a computer, extensive references, or tables.

Still another common approach is to ask students to recognize, rather than to actually carry out, the incorrect (or correct) application of principles learned in the course of instruction. In such a situation, the test task might be the description of an experimental design which an engineer has developed to test the effectiveness of a mechanical component, say an automobile door hinge, compared to other hinge designs. With this type of test item several true and false, multiple choice, or short answer essay questions may be asked of the student about the problem situation which is presented. These types of items can test for the learner's knowledge and ability to recognize correct or incorrect application of experimental design principles, to judge the magnitude of experimental results in terms of being reasonable or unreasonable, and to be alert to methodological errors. Such assessment tasks can be very demanding, can be administered in a much shorter time than would be needed to actually design an experiment or carry out some other complex performance, and can also be very informative about the degree of student learning. Perhaps an example will help.

An Example of a Course with Embedded Test Tasks

In an earlier chapter a course titled, "Hydrology and Sedimentology of Surface Mined Lands", was discussed briefly (Haan & Barfield, 1978). It was pointed out that the example problems which are found at the end of each of the course's six units define the functional competence and knowledge participants need to develop.to demonstrate the mastery of the course content. Over the three day period that the course is taught, the instructors present brief lectures.

Each lecture is focused around some common problem such as the task of designing stream channels for diversion of water from surface mining operations. The lectures and the printed materials, illustrations, tables, graphs, and related materials which are provided to participants in a single textbook are all closely interrelated. The purpose is to clearly illustrate: a) what the common problem is and the range of variation which may be expected in paramenters affecting the solution of the problem; b) the appropriate theoretical models, concepts, computational algorithms, and procedures basic to solving the class of problems; and c) the necessary adjustments, corrections, and modifications of models and procedures derived from theoretical and laboratory research to make them compatible with actual field conditions where more variables are operating in ways which cannot be controlled to the same degree as in a laboratory experiment. The printed instructional materials, as well as the lecture and problem solving activities of a practice nature which follow short lectures, are all designed to help engineers bring to bear the most appropriate knowledge, theory, and procedures for the design of hydraulic channels and storage structures. Students learn basic hydrologic theory and principles as well as many rules of thumb and specific procedures for making estimates about design specifications for such structures within certain probabilistic limits of maximum 24 hour storm rainfall and runoff, desirable safety factors in the performance of the structures, across various types of soil materials, and under differing slope conditions.

Much of the learning necessary to apply competently these engineering principles involves the use of complex nomographs and the use of statistical tables which list rainfall patterns and probabilities for various geographic regions. Soil types must be classified into various categories by which numerical values can be assigned to maximum permissible flow velocities to avoid channel erosion and sediment deposition downstream. Boundary properties of stream channels must be placed into various categories depending upon the type of vegetation in the channel or other cover and the resulting retardance to flow values. Many other variables including watershed ground cover, season of the year, and intended duration of the structures, must also be considered. For each variable there are one or more procedures by which to estimate a numerical value. There are other procedures for combining these values and equations to calculate design features of the stream channels and the storage structures. There are still other procedures by which to check independently the reasonableness of the design specifications calculated. It is important to double check and insure that the final design specifications are safe, cost efficient, and effective from the standpoint of their desired performance.

The course is worthwhile because it teaches specific procedures which are generally not available from any other source in such a well organized and integrated form. The textbook for the course is really a technical manual. It is specifically designed for the engineer to take back to the work setting in order that he or she may continue to use its tables, procedures, nomographs, and examples in the better solution of actual design problems.

Because of the complexity of the material and its proper application to real problems, the developers of the course embedded real problems as demonstrations in each section of the course. The first set of problems in each unit of instruction serves to illustrate how the course concepts and principles are applied. These demonstration problems are the main instructional tasks upon which instructors focus and in which participants engage to learn course concepts, skills, and procedures to solve particular types or classes of problems.

Immediately after these embedded tasks or problems, there is a detailed, point by point, explanation of how the problem can be solved. Often more than one method of solution is given. The textual materials provide a step by step illustration of how the course concepts and procedures are accurately used in problems sampled from the array of those encountered in the real world practice of surface mining activities. This allows an individual learner to study the background theoretical and empirical material presented in each chapter and then to work along with the demonstration problems in the actual application of these principles to real problems.

After the demonstration problems, a second set of parallel but different problems is presented. These are the embedded tests. They require the learner to demonstrate and transfer the application of course principles and concepts learned in the practice problems to another set of similar problems drawn from the same conceptual universe of real practice situations. During the course, participants are asked to solve these problems and allowed periods of from 1 to 2 hours to do so. Sometimes the problems are to be only partially completed or set up. Following this the instructors hand out completed solutions to the problems in order that these may be studied and used immediately (and in the future as well) by the engineer to check on the accuracy of his or her application of course concepts and procedures.

One other property of both the embedded practice problems and the embedded test problems needs mentioning. Both cumulate in scope and difficulty over the sequence of six chapters in the text. That is, the problems in the first parts of the course are small and consist of only part of the total task of designing the necessary hydrologic structures needed to divert and store runoff water for surface mining operations. Later chapters in the text require the use and integration of all of the preceeding material toward the design of complete drainage and storage systems. This is a strong feature of the course. It is difficult to integrate properly a large body of concepts and skills required for a competent performance in a complex area. There is no reason to believe that persons can easily learn to put together all of the parts of such complex processes unless they have been given practice and specifically instructed in how to do so (Bugelski, 1971; Gagne & Paradise, 1961; Gagne, 1967, 1977; Snelbecker, 1974).

Reasonable Expectations for Achievement Within a Short Course

What is it reasonable to expect in the way of learning outcomes from a three day participation in this continuing education "Hydrology and Sedimentology of Surface Mined Lands" short course? Is it possible that enrollees will become facile with all of the complex theories, concepts, and procedures encountered in the course at the end of the three days? In fact, it is not. About all that can be hoped for is that the participants will:

a) become highly motivated to continue study and use of the course textbook or manual because they see how useful it can be to them in improving their performance in their daily work tasks.

b) become familiar with the main ideas, concepts, procedures, algorithms, and the adjustments which must be made in these to accomodate differing soil types, rainfall patterns, surface cover, and so forth.

c) know how to use the many procedures, nomographs, tables, charts, and models presented in the manual in intelligent ways to produce reasonable solutions. Much of this is knowing when, why, and how to use a particular approach to solve a particular design problem, knowing what the reasonable values should be for certain types of problems, and knowing how to check on the accuracy of one's calculations by independent means.

The content and skills of the course have considerable utility for the participants, particularly if they are involved in surface mining engineering. Some weeks or months after the course is completed participants may become more skilled in the use of the procedures through application of course content on the job. Many highly technical continuing education courses share this characteristic.

Practical Problems in Using Complex Embedded Test Tasks in Short Courses

Suppose now, that one wishes to evaluate the learning outcomes for the course at the end of the three days. How might this be accomplished? One way might be to require each participant to complete individually and fully each of the many embedded tests or problems at the end of each chapter or unit of study. These could be collected, corrected, and an individual conference could be scheduled with each student to report their progress and to instruct the student in any areas needing more attention. The final, and very complex, cumulative problem tasks provided in the Hahn and Barfield (1978) text are excellent assessment tasks. Their completion requires the proper use and integration of all the prior learned information and skills. As sound as this procedure is with respect to instruction and the assessment of learning outcomes of the participants, there is a serious problem with this approach.

The problem is that it would take much longer than the available time to carry out such an assessment, perhaps six to nine days on the average rather than the three which are available. The persons who come to short courses often do so only by making sacrifices within an already very demanding schedule. Furthermore, the instructors need to spend most of their time and effort in instructional activity, not in administering a massive test which takes days to complete and even more days to score. Moreover, many participants may have attended for purposes of acquiring an overview to the content and methods of the course, not to become proficient in the actual design of drainage and storage systems. In addition, most participants would probably perform poorly on such a comprehensive test. Fully competent performance on these types of complex problems is not expected as an immediate outcome for the three day short course. The limited time available also makes the completion of massive amounts of homework or major test problems by participants not feasible. The correction of massive amounts of completed homework by instructors is also not feasible. About all that can be done to work through some good illustrative demonstration problems and to complete parts of "homework" problems under the supervision and assistance of the course instructors. How then, could one design a test to determine if the learning outcomes for the course had been achieved? In what ways might test items or tasks be developed which would be effective measures of performance in key areas, but which would be able to be administered and scored in a matter of a few minutes rather than several hours or days?

Abbreviated Embedded Test Tasks - An Illustration

Appendix B contains a sample multiple choice test which was designed for the hydraulics of open channels section of the "Hydrology and Sedimentology of Surface Mined Lands" course. This is one of the six units of instruction in the three day course. The total test length for this unit is eleven items. The time required to complete the test is approximately ten minutes. Yet the test items assess most of what is of interest with respect to the achievement of specific learning outcomes at the end of the unit of instruction. Similar tests of from five to ten items can be prepared for each of the other five units of instruction. These can be used at the end of each unit of instruction in the course. A time limit of ten minutes can be set for each test. Participants can complete the test immediately after a unit has been taught. Because of the objective nature of scoring, the participants' responses can be corrected immediately with a scoring key, requiring no more than ten to fifteen seconds per participant's paper. The results of the performance on the test by participants on an individual basis and for the whole group can be reported to individual participants within minutes of the test administration. (More details about how to carry out this procedure are presented in Chapter 13.) Common points of misunderstanding can be noted by instructors and remediation or correction attempted in subsequent instruction and dialogue. The results from all such short embedded tests can be added together across the six units of the course and reported as an indication of an end of course learning outcome for particular students. Results can be summed and averaged across students to make inferences about the success of the course for purposes of formative or summative evaluation.

In addition, duplicate or parallel items can be written for each of the items in the embedded tests for each of the units. These can then be assembled into a pre-test, a post test, or a delayed post test to serve other purposes.

The point to all of this is that embedded test tasks are often excellent learning assessment devices because of their directly parallel structure to the instructional activities and tasks and because of their practical nature. Yet they are difficult to administer, score, and use to instruct students about errors in learning and areas needing further attention, especially in courses operating under severe time limitations. Because of these factors, short objective type test items, consisting of abbreviated performance tasks, similar to the eleven items presented in Appendix B ought to be designed and used more often. How may such short but powerful objective type test items be developed? Perhaps examination of the purposes and properties of the eleven items displayed in Appendix B can help answer this question. Not only is the example test provided in the Appendix, but detailed instructions about how to prepare abbreviated test tasks in general are provided. Additional information about how to design and use efficient abbreviated test tasks is also found in Chapters 10 and 13.

Purposes and Properties of Abbreviated Embedded Test Tasks

Appended to the set of eleven test items in Appendix B are a group of figures and tables. These are presented to the students along with the test items. The hydrology and sedimentology course teaches the proper approach and solution of problems through the use of many tables, nomographs, and similar procedures. An important part of the task for the student is to learn how to use such material as well as to discriminate from among the many tables and figures those which provide the information needed for the solution of a particular problem. By having this series of graphs, figures, and tables in one place as an appendix to the test items, the student must discriminate among the multiple displays to select the appropriate ones for the solution of a problem part presented in an individual test item. In addition, the student must know how to enter the table or graph and retrieve the information needed. Thus, the test items assess two skill areas which are key objectives in the course. Each item simulates part of a real problem situation.

The items for the most part do not require computation. Items one through three test for basid knowledge of basic properties and relationships required to solve this class of problems. Items four through eight are all constructed around one common problem situation concerned with the design of a particular open channel given certain performance specifications, soil type, cross section, and slope. Items four and five test for knowledge of how to enter the correct tables and extract the correct values for two design variables given specified problem conditions. Item six tests for the student's knowledge of an estimation procedure and the use of the procedure to double check on the initial values obtained for channel specifications. Item seven tests for the student's ability to recognize the reasonableness of a result based upon a short cut estimation procedure. Item eight is the first item to actually require any computation. Its correct solution requires the individual to use information presented in the original problem statement preceding item four and the additional information given in item six. It requires the individual to actually calculate the top width, bottom width, and the depth of the channel allowing the necessary freeboard. To answer this question properly, one must know which variables to attend to, which computational algorithms to use, and again, be aware of the range of resulting values which are reasonable given the problem characteristics. Items 9 through 11 are another series of similar questions written around a common problem situation. These three items test for knowledge of procedure, use of rules of thumb to modify models and computational procedures, and checks on estimation procedures. The skills tested for by these three items are similar to those tested for in items four through eight, but the problem characteristics have changed.

Collectively the eleven items test, quite well, the level of competence achieved in the basics of solving these types of problems. Yet the time required for this testing is small because of the manner in which the tasks are presented in the items. The emphasis is not upon calculation, but upon knowledge of procedure and proper practice, although actual computational skill in determination of design specifications is also tested. In addition, there is a hierarchy of difficulty in the test items. The first two items test for basic concepts and knowledge. Later items test for knowledge and skill in the use of proper procedure. The last items require integration of the knowledge and skill required in all of the earlier items and actual computation of design specifications given different characteristics in the actual problem situation. Consequently, the results of students' performances across items tells something about what parts of the intended learning outcomes they have achieved well or not so well. This is useful information for individual learners in order that they may correct or further develop their knowledge and skill in needed areas. It is also information very useful in the summative evaluation of the effectiveness of a given course when the pattern of correct and incorrect responses is examined across all students. If there are problems in teaching certain procedures or methods in the course of instruction, these are likely to show up in the test results on specific items or clusters of items.

Advantages of Multiple Choice Items as Embedded Test Tasks

It should be clear from the example items in Appendix B and the accompanying text that multiple choice items can be written to test for very complex and high level skills as well as for recall of factual information. Many times multiple choice test items are written only at the factual information recall level. A test composed of such items does not provide information on higher level capabilities such as skill in formulation of a problem, using computational procedures correctly, applying estimation procedures, and integrating information from several different sources to make informed decisions about how to proceed in the solution of a given problem. Tests composed of items similar to those listed in the sample test in Appendix B are much more valid for such purposes than are tests composed of more typical factual recall items.

Tests similar to the one displayed in Appendix B are also more efficient than other types of longer embedded performance tasks. The reasons for this are that they can be completed in much less time, can be scored very rapidly with a standard scoring key, the results can be tabulated and presented to the learner immediately after the test is completed (see Chapter 13), and, if properly constructed, they are good approximations to more complex, longer duration performance tasks and work samples.

Precautions in Developing Multiple Choice and Other Objective Test Items

When developing tests similar to the one displayed in Appendix B there are a number of precautions which should be observed. First, it is best to develop the individual items from full blown, actual problem situations or tasks. That is, the actual complex and very time consuming problems which are usually assigned as homework problems or which are used as laboratory or demonstration activities, ought to be the basis of the individual and abbreviated multiple choice test items which are generated. One should start with the real problem or problems representative of the performance domain which is being taught. Then one should work out these actual problems in full. Then one should ask, "In what ways might I break the solution of this problem down into individual test items of a multiple choice format? The task is to capture all the critical aspects of the decisions and judgments which persons must make in the solution of real problems in all of their complexity with a small number of test items. Different parts of this decision making and judgment can often be captured in different multiple choice test items written about one common problem situation. Certain information can be given so that the person must not necessarily do all the actual work required to solve a real problem of the type under consideration, but so that he or she must discriminate the relevant variables, apply knowledge of procedural rules, select appropriate models, modify computational procedures according to problem characteristics, and so forth. Beginning with a real and complex problem, typical of the type encountered in the real domain of work, helps insure that the test items written to assess knowledge and skill about specific areas of procedure will be significantly related to intended learning outcomes. Collectively the items should be a reasonable test of the breadth and depth of the person's skill. Basing test items upon real problems also insures that inferences made about the learner's achievement from test scores will be reasonably valid.

If one begins the other way, by writing only those multiple choice test items one happens to think about, without referencing each item in some actual complex problem, one is likely to generate many factual recall and simple information items. In addition, the total test comprised of the assembled items will often not be very broad or test for depth of skill across areas of complex performance. Just as is the case in the selection of tasks and activities by which to instruct persons and provide them with practice in learning complex performances, so too in test development it is best to begin with a typical set of complex problems of the type likely to be encountered by the practicing engineer. Once these problem types have been selected, it is possible to break them down into smaller components for purposes of instruction or for testing of student achievement. Chapter 10 provides detailed procedures for insuring that both test items and instructional tasks are sampled from the full performance domain to insure validity for each. Consequently, no more will be said about this methodology at this point.

Importance of Item Independence

Another caution is necessary. For methodological reasons it is important that the individual test items be logically independent from one another with respect to a correct answer on one item being required for the correct answering of other questions. That is, one does not wish to prevent the student from being able to demonstrate knowledge or skill on subsequent test items because he or she obtained a wrong value in response to an earlier item. Consequently, later items should not require correct responses to earlier items as a condition of their correct solution.

This does not mean that multiple sets of items may not be written about some common problem situation. Nor does it mean that information needed to solve a later problem in a series of such items cannot accumulate in subsequent items. It does mean that when such information accumulates it should do so in the item stems, and not in the options which are presented as alternatives from which the student is to choose. Items four through eight in the sample test in Appendix B illustrate this point nicely. All four items depend for their correct solution on information presented about a given problem. This information appears in a stem which preceeds all of the items. This common stem appears prior to item four. Students are told they will need to use this information for the next four items. To help the student remember that this block of items goes together, all four questions and the common item stem are enclosed in a bracket on the left margin of the test booklet. Furthermore, information presented in the stem of item six is required to correctly answer item seven. However, in no case does any item require that a student answer any previous item correctly if he or she is to obtain the correct answer to a subsequent item.

It is important to maintain such item independence. If an entire series of items depends upon the correct answer to an earlier item, all of the remaining series will be incorrectly answered if an error has occurred in response to that first item. The subsequent test items are not valid indicators of what the individual knows or can do under such a restriction. The value of any individual test item is to assess some particular knowledge or skill. Through multiple items designed to assess multiple areas of competence at various levels of skill, it is possible to make strong inferences about the extent of learning of the individual based on the total test score. If it is of interest to determine if the student can correctly work through a complex series of procedures, where in real life the correctness of the next step depends upon the accuracy of the previous steps, this can also be tested. However, it should be tested in a separate and more complex item. Item eight as well as items nine through eleven are this type of item on the sample test in Appendix B. This complex type of learning outcome is best tested for by a performance task which simulates a complete real world problem.

Generalization of Item Construction Procedures to Other Test Formats

It should be clear that one does not need to use multiple choice items to achieve the objective of having a short but powerful test by which to assess the degree of learning of students. Each of the multiple choice items shown in the sample test in Appendix B can be used as a constructed response test item. That is, each stem for each multiple choice item may be used without any of the distractors and the correct answer which presently comprises the options from which the student is to select. If the options are omitted the student must simply construct the correct answer. Provided the item stems are written to call for particular types of judgment, skill, concepts, or application of procedure, the constructed answer can be short and can also be easily scored in an objective manner. There is also an advantage in using the constructed response format in that there is less possibility of guessing producing a "correct" response. On a four option multiple choice test consisting of n items, by chance alone a score of .25n will occur across persons on trial administrations. One compensates for this by insuring that there are enough items to test a broad range of knowledge and skill and by setting an acceptable level of performance well above the chance level. Various methods exist for correcting for guessing on multiple choice tests. These involve subtracting more than one unit value of incorrectly marked items from the total number of items marked correctly, but not doing so for items the student left blank. Generally, however, it is better to include enough items on the test and to set a criterion well above the chance level than to correct total test scores for "guessing" when using multiple choice formats.

If the class is small and there is time to administer a test similar to the sample test in Appendix B it is probably better to do so as a constructed response test. The scoring of these tests takes more time and intelligence. However, if the class is large, and time is at a premium, it is probably better to use the multiple choice format. This is because students may place their answers on a standard answer sheet which may be scored immediately with a hand scoring key or by machine. This is a very rapid and accurate way to proceed and insures that students can receive the results of their knowledge assessment right away, a basic requirement for good use of tests for instructional purposes.

Whether the test format is multiple choice or short answer constructed response, the most important consideration is to have well developed test tasks in the form of questions or item stems which have been sampled properly from the performance domain being tested and which also have been designed to test for specific skills or knowledge. Tests designed to meet these specifications make excellent embedded test tasks. They may be interspersed as short quizzes at key points in the instructional sequence. Their use takes little time and reveals a great deal to the instructor and the students about the achievement of particular learning outcomes. For additional examples of how to develop tests with these characteristics, the reader should turn to Appendix B and Chapters 10 and 13. Well designed abbreviated test tasks are particularly important in short courses where there is inadequate time to use the assignment of homework problems and laboratory activities which are the embedded test tasks most frequently used in traditional, long term courses in technical fields.

Generalization of Item Construction Procedures to Pre-, Post, and Delayed Post Test Construction

Most of the details and advice presented in this chapter concerning how to go about selecting and developing good embedded test tasks and items also applies to the construction of pre-test, post test, and delayed post test items. To the extent that test items and tasks are developed along with instructional tasks and learning activities, as both are sampled from the common domain of expected performance, the test items will be better than if otherwise produced. This is because test items are actually specific tasks which are reserved for careful observation of student performance. From the observation of student performance on this sample of tasks, the instructor makes inferences about the degree of student learning of specific knowledge or skill. Any one test item is not sufficient to making inferences across the entire domain of performance. Rather, individual items should be constructed to determine the student's level of skill or knowledge in specific critical aspects of the performance. Test items must be properly representative of the range of knowledge and skill required for effective performance of the complex activity being taught. Student performance across a number of well developed items of this type can be valid measures of the degree of student learning and the success of instruction.

Because test tasks or items are so closely related to the topics, materials, and activities of instruction, it is best deliberately to develop test items at the same time one is developing instructional tasks and activities, a point repeatedly made throughout this book and a procedure described in detail in Chapter 10. It is also best to use these assessment tasks during the course of instruction as embedded assessment procedures to inform both learner and instructor of the progress, accomplishments, and problems of individual persons in order that errors may be corrected and areas not in need of further instruction be omitted from further instructional activity. If such procedures are followed, one develops a large pool of assessment tasks and items. Assemblages of these test tasks or items may be prepared and used as time efficient and yet highly valid pre-tests, post tests, and delayed post tests.

Conclusions

Traditional forms of embedded tests including quizzes and homework problems are common instructional methods. It is possible to abbreviate complex performance tasks into short and efficient test items by which to make sound inferences about how much and what students have learned following units of instruction. It is best to develop test tasks at the same time the instructional activities are being developed for teaching the course. Both test tasks and instructional tasks should be developed from complete and typical complex problems the engineer will face in the work setting. This helps to insure that the test items which are developed, as well as the practice problems which are assigned, will consist of a range of difficulties and include a hierarchy of concepts and skills required for effective performance. Care should be taken to insure abbreviated tests include items of different difficulty and that collectively the test is a good assessment of the range of content knowledge and skill which is desired.

The procedures which apply to the development of good embedded test items also apply to the development of test items for pre-tests, post tests, and delayed post tests. The procedures also apply to tests of different formats, including multiple choice, essay, short answer, and problem completion types of items. The main purposes of abbreviated embedded tests are to help diagnose learner achievement and learning problems, provide information on course effectiveness usually in a formative way, and serve as a part of the learning experiences and activities of students.

Although embedded tests are most useful in providing information about how much and what students have learned or not learned immediately following a unit of instruction, the information they provide can also be used to make inferences about the achievement of students in meeting the overall learning outcomes set as objectives for the course. This can be done by keeping a record of performance of individual students across all of the embedded tasks for each unit of work. Collectively this record can be used to draw a learning profile for individuals and groups. Under typical situations the profiles of students may be expected to improve with time and toward the end of the course as more information, concepts, and skills are mastered.

Information of this type is also useful in determining which parts of the course are most effective and which parts need to be improved. Programs or courses of instruction are almost never totally good or bad. They are usually a mixture of effective and less effective learning activities and instructional methodologies. By being dispersed throughout the course, administered after each unit of instruction, and because they are specific to well defined sections of courses and particular concepts and skills within these sections, embedded tests provide diagnostic information. This information is useful to the students in the course who learn right away what they do and do not understand. It is useful to the course instructors in immediate modification of instruction to better meet the needs of students. In addition, it provides insight into which instructional activities and teaching methods are working well and which ones need to be improved. For all these reasons, embedded learning assessment tasks are very important tools, especially in formative evaluation activities.

Chapter 8

POST TESTS, THEIR PURPOSES AND USES

No matter how adequate the embedded test tasks may be in a course or unit of instruction, typically some indication of students' end of course achievement is needed. This is a summative evaluation function which serves to inform the learner, the instructor, and appropriate others (such as employers, professional certification agencies, and governing boards for continuing education units), the amount of student learning at a particular point in time.

Cumulative Nature of Post Tests

A post test should be cumulative across the various levels and sequences of skills and knowledge instructed in the course. That is, there should be items or tasks which test for knowledge and skill from all portions of the course. Items should be included requiring: a) recall of simple information, b) recognition of correct principles, c) use of appropriate concepts and procedures, d) formulation of problem situations in terms of specific and appropriate sequences of activity, and e) correct technical solution of complex problems. If the post test measures only recall of facts or recognition of procedure, nothing can be inferred from the test scores other than students' capabilities in these areas. Consequently, care must be used to assemble post tests which include a proper range of tasks across the major levels of skill and knowledge required for effective performance in the area being instructed.

Suppose one is teaching a short course on the use of microprocessors in automated control of industrial machine processes. Now suppose a post test is developed. The test items require *naming* the typical components of such systems, *recognizing* proper and improper logical steps in programming such systems, and *naming* the producers of certain types of equipment needed to implement microprocessing controls. After the test is administered, what types of inferences may be properly drawn from examination of students' performance scores? Assuming that the test items are all valid and properly constructed, inferences can be made *only* about the ability of students to *name* the components of such systems, *recognize* proper and improper logical steps in programming, and *recall* the names of equipment producers. What is being tested is naming, recognizing, and recalling certain facts, principles, and names of companies. In no way can performance on such a test be construed as demonstrating mastery of actually planning, assembling, and integrating microprocessing equipment into industrial machining production processes. This test is perfectly appropriate if the intended outcomes of the course are the *naming*, *recognizing*, and *recalling* goals. However, if the objective for the course is the performance capability of integrating microprocessing equipment into industrial machine production processes, the post test is very inadequate as an assessment task.

One way around this problem is to select test tasks to represent the full range of the instructional tasks developed for the course, a prodedure suggested in the last chapter on embedded test tasks and described more fully in Chapter 10. If one has a range of items parallel to the various levels of instructional tasks in terms of knowledge and skill required, a good post test can easily be assembled. But if one has the embedded test tasks and one has used them all along during the course of instruction, why develop and use a post test at all? There are two reasons why a post test makes sense even if embedded test tasks have been used throughout the course of instruction.

Revealing Cumulated Learning

A simple summation of a student's performance across all of the embedded tasks may not be an accurate assessment of his or her competence achieved at the end of the course. The reason for this is that embedded tasks are useful primarily for providing the student and the instructor with information about what is being learned well and what is not during instruction. The results of individual embedded test tasks often reveal weakness in a student's learning, flaws in the instructional methods, or inadequacies in the instructor's ability in teaching particular skills, procedures, or concepts. When these problem areas are identified they are usually corrected. When embedded test tasks or items are used in this manner they contribute greatly to effective teaching and mastery of key course skills and knowledge by students. This is a formative evaluation function. Information is provided by which future instructional activities of the learner and the instructor are modified to achieve mastery of course content. Thus, students who initially do poorly on specific test tasks or items embedded in the sequence of instruction, should master the areas in which they are having difficulty by the end of the course. There is a great deal of empirical research which indicates this is the typical pattern in courses where embedded test tasks and their results are used in a formative way. The Personalized System of Instruction Method (PSI), sometimes called the Keller plan, operates precisely this way. PSI has proven extremely effective for the instruction of engineering and other technical and scientific courses (Cleaver, 1976; Ericksen, 1974; Heimback, 1979; Kulik & Kulik, 1975; Kulik, Kulik, & Cohen, 1979).

For this reason, a simple summation of all of the embedded test tasks performances of students is not accurate. It provides too low an estimate of students' exit level learning. Because of this, nearly all PSI and similar mastery learning approaches to instruction typically provide some final and comprehensive test of performance. These tests usually consist of items sampled from across the entire course of instruction. The items are parallel but not identical to the items used in the embedded tests during the course of instruction. This insures that students are required to learn general skills and principles and not simply the array of correct responses to a particular set of items, a possibility if the same items were used on both the embedded test tasks and the final examination or post test. There is much research on the effectiveness of instruction which makes use of embedded test tasks in a formative way coupled with the use of summative post tests. Where the focus of instruction is upon mastery of complex skills and performance capability these testing procedures are very effective (Gagne, 1962; Grogan, 1979; Kulik & Kulik; Kulik, et al., 1979).

Measuring Functional Integration of Knowledge and Skill

The second reason for use of post tests is to provide an assessment procedure by which to test for the student's ability to integrate and use wisely all the component knowledges and skills which have been taught in a course. On the basis of individual tests of separate components of the total performance, it is not possible to infer that persons have learned to put all of these components together into a skilled performance involving the solution of a complex problem. If one wishes to test for skilled performance in integration and wise use of the range of skill and knowledge required for solving problems of particular types, one needs to develop test items which have this particular requirement for their correct completion. Again, increasingly complex embedded test tasks of the type included in the Hydrology and Sedimentology course referenced earlier (Haan & Barfield, 1978) and similar to the sample test shown for one unit of that course (see Appendix B), can do much to assess the student's ability to perform the integration of particulars. Good embedded test tasks cumulate across the course from test tasks of simple and particular knowledge and skill to very complex test tasks requiring discrimination, judgment, integration, and wise use of a large amount of facts, concepts, procedures, principles, and theories. Performance-based engineering education approaches are designed in this manner. A series of successful project activities and assessments of complex performance capabilities are scheduled throughout a student's ppogram. The complexity of these performance tasks and projects increases systematically as the student progresses (Grogan, 1979). If this is so, why would a post test which tests for similar ability to integrate and use course knowledge and skill wisely and well be needed?

The answer is similar to an earlier answer concerning the use of post tests. Even complex embedded test tasks are more intended for practice and instructional diagnosis of what students need to learn and what they have already learned at particular points in the course than they are for summative evaluation statements about the overall amount of learning resulting from the course. Even the most demanding and complex embedded test tasks are typically practice situations. They are usually administered in situations where the student may ask for and properly receive assistance if he or she has difficulty on any point in the problem. This is appropriate because the task is seen as a way to diagnose what the student can or cannot do in the interest of achieving mastery of the course content. Both the instructor and the student need to know what must still be learned to perform some complex task correctly and efficiently. Collectively the record of an individual's performance on these types of embedded test tasks can reveal much about the student's rate of progress through the course. Sometimes it becomes clear that a student is having great difficulty and progressing too slowly. If a student needs too much time and extra assistance to master material other students can master much more quickly, it becomes a serious drain on limited time and teaching resources. Likewise, a record of individual student performance on embedded test tasks also reveals which students are highly competent and rapid learners of the course content. Yet when all is said and done, it is usually desirable to have an independent estimate of the student's learning at the end of a course to sum up the overall degree of learning. The results of embedded test tasks and these final assessment tasks may be combined and used as an overall estimate of the learning resulting from a course. The reasons for needing this summative data have been mentioned before. They include the needs of the learner, the instructor, the employer, the professional credentialling agency and others legitimately concerned with the degree of learning achieved by the professional engineer at the conclusion of a period of training in a continuing education course of whatever variety. More will be said about this in Chapter 13.

Summary of Main Features of Post Tests

It is apparent that good post tests for continuing education courses must have three main features. First, they should be representative of the breadth and various levels of course knowledge and skills. Second, they should test the ability of students to integrate all of this knowledge and skill wisely and with technical accuracy in the solution of complex problems. Third, they should be short, require only a small amount of time to complete, and permit relatively quick and objective scoring procedures. It is for the third reason that the testing of complex skills by the military, the government, and many industries has come to rely upon multiple choice and similar short answer objective items. Tests developed in this format, which follow the procedures laid down in this and the other chapters, are likely to be good estimates of the degree of learning outcomes resulting from continuing education courses. They are also likely to be efficient in terms of the time required for administration and scoring. In addition, if used properly, they are likely to be appreciated, rather than objected to, by enrollees in continuing education courses. Persons who are learning complex skills and procedures usually welcome the chance to demonstrate their newly developed competencies. A large majority of the professional engineers surveyed in a number of continuing education courses at the University of Kentucky and elsewhere have indicated they have no objection to being tested (Ferry, 1979; Moss et al., 1978).

Adequate versus Ultimate Summative Evaluation

Post tests are needed to provide independent and efficient assessments of the degree of learning which occurs at the end of a course. These assessments are only estimates. The degree to which the estimates are valid and accurate depends upon how well the tests have been constructed. There is no substitute for the assessment of complex performance and the products which result from that performance in actual work settings. The ultimate summative evaluation of a bridge design is how well an actual bridge built to the design specifications performs over a long period of time, perhaps 80 years or more. Yet it is not possible to test the bridge design by waiting 10, 20, or 80 years to see how well it actually performs. Rather, one must take what one knows about bridge design and construction generally, design small models and simulations which test aspects of the proposed bridge design, and finally construct inferences on the basis of these simulations and the general fund of knowledge about how to actually build bridges. Then bridges are built, even though the ultimate evaluation has not occurred.

So it is also in the design of courses to teach persons complex skills and performance capabilities. One cannot wait and see how well the performance capability has been learned in a long term sense by course participants before teaching the course again anymore than bridge builders can wait for the ultimate full performance test of a bridge design before constructing additional bridges. In both cases there is too little time to perform the total and complete summative evaluation.

Even the supervisors of professional engineers sample only parts of the engineer's actual performance for close examination. There is simply not time to observe the entire performance. In addition, the supervisor samples, for detailed examination, only some of the products of the engineer's work. The beginning engineer is watched more closely, the highly experienced and skilled engineer very little. It is even less practical for the continuing education instructor, or the director of a continuing education course, to monitor closely the actual on-the-job performance of engineers who have completed training in particular short courses and other types of continuing education experiences. All that can be done is to construct reasonable approximations to some of the more critical features of performance of some complex procedure. These are usually abbreviated simulations of aspects of real tasks and problems which will be faced by the engineer in his or her work activity. These tasks can be test items similar to those displayed in Appendix B or some other form of more practical examination as is often given in laboratory or project assignments in technical courses. From these types of assessment tasks it is possible to make inferences about how well participants are learning the skills and knowledge which are the intended learning outcomes for specific continuing education courses.

Chapter 9

DELAYED POST TESTS AND OTHER ASSESSMENTS OF LEARNING OUTCOMES

No matter how excellent are the results of the initial tests of a new bridge design which uses new construction materials and methods, the persons involved are keenly interested in the actual performance of the bridge once it is in service. Consequently they observe it closely, note any problems, and modify designs for future bridges. There is a parallel for persons who develop and operate continuing education courses. No matter how popular the course, or how positive the initial summative evaluation, it pays to monitor the performance of persons who have completed the course to learn more about the need for the course, the degree to which course content and principles are being used in sound and proper ways, and the general value of the course to the engineering population it is designed to serve. This constitutes the first goal for delayed post tests or other delayed assessment procedures. The information from these follow-up assessments reveals not only what the participants have learned, but also whether they are able and willing to put this knowledge and skill to use in their daily work activity.

Complex Skills Improve with Time and Experience

There is another reason for the use of delayed post tests or other types of follow up assessment procedures.

Many of the outcomes for continuing education courses in engineering are very complex performance capabilities. These capabilities are based upon the proper conceptualization of many facts, relationships, and concepts; and upon the ability to apply many principles and theories to a wide range of practical situations. Since these situations are very diverse, many times modifications must be made in procedures, principles, and models if particular problems are to be solved. No one course can offer the range of problem situations needed to fully instruct participants. The high levels of competence in applying course content and knowledge to the entire range of situations likely to be encountered in the real work setting come only with much appropriate experience. The learning activities included in any course are also only partial samples of what must really be experienced and understood if performance in the domain under consideration is to become highly expert. In such situations the usual goal is to make the student very familiar with the basic features of what a good performance is across a finite number of realistic situations. In addition, one usually also tries to teach something of the background, theory, principles, and varialbles which help the engineer understand better why certain procedures work well and others are not particularly suited to the solution of certain types of problems.

It has long been known that when skillful performance of complex tasks is the goal of instruction, quality of the performance usually increases after formal instruction is completed and the person has returned to the job or continued to pursue study in a related area. Simple facts and recall of information decresse rapidly with time after instruction. Yet, retention of principles, techniques, and skills in complex performance areas typically increases and becomes even more polished long after formal instruction has ceased, provided they are used.

Tyler (1934) performed some of the early empirical studies which revealed this principle. He found that college students in a zoology course forgot 77 percent of the factual material, such as naming parts of animals, a year after completion of the course. However, there was no loss in skill in applying principles or rules learned in the course to new situations not encountered in the course, even a year after the completion of the course. In fact, Tyler's studies showed a 25 percent gain in the skill of properly interpreting new experimental designs in new areas a year after the course had been completed (Gage & Berliner, 1975, p. 143).

The same results have been observed many times by other researchers. Complex skills, principles, and procedural methods are difficult to learn. One cannot simply "look up" these skills and know how to use them. They come only with much effort and guided practice; and they get better with time and experience after instruction is completed in the formal sense (Cole, 1972; Gagne, 1968, 1965, 1974). Reading is one example of a complex skill with exactly these types of properties. One cannot "look up" how to read. It is a complex performance consisting of many sub-skills which must be integrated into an efficient and smooth performance. These types of complex learnings have been referred to as "process skills" because they are generalizable ways of processing information and solving problems. Once learned they are very resistant to forgetting or what psychologists call extinction. They typically remain life long, and are used continuously. Like good wine, they improve with time (Cole, 1972).

Measuring Growth of "Process" Skills

Because the skills being taught in many continuing engineering education courses are of the "process" variety, it is important to determine if participants who have completed the course some time ago are growing in the skills they acquired initially during the course. If they are not, it probably means they are not using the information and skill offered in the course in their work. This may be a good indication that the course is not particularly needed or relevant. Of course, this may be true for some courses and not others. Again, it must be remembered that some engineers and scientists enter even very technical and specialized courses not because of the desire to use the information learned in their daily work activity, but simply because they are curious or wish to become more broadly informed. However, if it is presumed that the main reason for a continuing education course is to upgrade specific knowledge and skill related to increased competence in some engineering activity specialty, failure to use concepts and skills acquired in a course by large numbers of participants over replications of a course is damning. If the course is central and vital to better performance, participants should at least not forget or perform more poorly on a delayed post test or other assessment procedure compared to a post test. Ideally and practically, they should improve in their performance, particularly if they are applying course principles and procedures frequently in their work activity. It must also be remembered that participants will apply selectively that which they have learned. All of the knowledge and skill acquired in a course may not be routinely used on the job. However, these portions of the course which are relevant to the work activity of the engineer should come to be learned very well.

Now that we have examined two reasons for the use of delayed post tests or some other type of assessment procedure, let us consider other reasons for such assessments.

Purposes for Assessing Delayed Learning Outcomes

Formal administration of tests developed and used earlier as post tests is one method of determining how much of course content and skills is being used and retained by past course participants. If such a procedure is carried out, one would not usually test all participants from past course sessions. Rather, one would sample from among those persons, mail out the test, have persons complete the test, and return it for scoring. Generally it would be of little interest to the practicing engineer and his or her employer to participate in such delayed post testing as a means of assessing the individual's competence in complex performance areas. As was noted earlier, it is the daily observation of the engineer's performance and the products of the performance which constitute the practical evaluation carried out by the engineer, his or her colleagues, and the employer and clients for whom the engineer works. This practical evaluation and the tacit understanding which it produces among all of these groups is the most meaningful and significant delayed post test evaluation. If a course is found wanting by this tacit evaluation procedure among these groups, it will not be subscribed in the future and it is, in fact, functionally judged to be inappropriate to needs of practitioners.

Why then, would it be advisable to administer a formal delayed post test? Why would participants of past courses be willing to engage in such an activity? The answers to these questions have to do with the need to conduct formal summative evaluation of courses and their effectiveness by those who design and teach them. The central issue is the evaluation of courses, not the evaluation of persons. A formal testing of the knowledge and skill of a sampling of past course participants can tell the course instructors and developers if the content and procedures taught in the course, are, indeed, being practiced and learned to higher levels of competence by participants after returning to the job. If this outcome is a goal for the course, it ought to be measured. If the outcome is achieved as indicated by an improvement in the scores of participants on a delayed post test, this information can be very useful. It can provide some indication of about how much additional practice is required for participants to become facile with the content and skills of a given course. This information can be made public to potential future participants and employers. In that event both may have a more reasonable expectation about what to expect as an immediate learning outcome for a particular short course or similar experience. The information also communicates what may be required in the way of on-the-job learning following the course experience to help the participant become competent to high levels of expertise in a given skill area. This approach is especially important in situations where there is a great amount of complex material to be learned in a short time.

The "Hydrology and Sedimentology of Surface Mined Lands" course is a good example (Haan & Barfield, 1978). The material in this course is so complex that all that can be expected as a reasonable immediate outcome for the three day short course is that participants will understand how to approach the design of drainage systems and storage structures using the latest thinking, models, and techniques. Thorough familarity with the appropriate methods and procedures is the immediate goal. In addition, the course textbook is really a technical manual which contains all of the information, models, procedures, tables, computational algorithms, and rules of thumb typically needed to design any drainage and storage system for any surface mining operation on any type of topography, soil type, and climatic condition. Therefore, a major outcome intended for the short course, as an immediate objective, is great facility in knowing how to use the manual. This involves knowing how to locate appropriate information from charts and tables, knowing how to set up a problem given certain conditions and ranges of values in parameters such as rainfall patterns, soil types, slopes, local ground dover, and related matters. At the end of the course, all participants should be able to demonstrate high levels of competence in the correct use of the manual for a range of sample problems.

These sample problems are the test tasks which are embedded throughout and the post test tasks found at the end of the course. Yet, ability to perform well on these tasks does not insure continued growth and facility in actually designing good drainage, diversion, and storage systems for surface run-off in surface mining operations. If the initial course objectives are achieved the participants are equipped to begin to improve their actual designs in this area through the use of many new conceptual and procedural tools presented in the three day short course. Actual proper use of these tools will occur only if the participants return to their work setting and continue to design such systems and apply the tools and skills they have initially acquired.

Work Samples as Alternatives to Formal Testing

Another alternative to the administration of a delayed post test to participants after the completion of a short course is evaluation of actual work samples. Enough time should have elapsed to allow for the participants to actually have engaged in the repeated application of the skills and knowledge learned in the course. Actual work samples from on the job performance are collected and evaluated. For example, a random sample of past course participants can be asked to submit a recent actual drainage and storage design for a surface mining operation which they had prepared. These actual products of engineers' performances can be evaluated and scored much the same way a laboratory activity of students is scored. Persons who teach the course can determine to what extent the concepts and skills in the course and the technical manual are being used properly and efficiently. Common errors or misunderstandings can be noted by instructors. This procedure constitutes the best possible delayed post test assessment of the participants' ability to actually use the course content and skills in ways to improve practice. However, the procedure is time consuming and difficult. It might take as long as a day to "grade" and evaluate the design of any one engineer for a given complex problem. Certainly only a few such thorough evaluations could be carried out in terms of available time. The value of such activity would be primarily for a stronger summative evaluation about the effectiveness of a course than could be obtained from an end of course post test alone. Again the interest in carrying out such delayed post testing is for the evaluation of courses, not persons.

Advantages of Abbreviated Test Tasks for Delayed Assessment

A more efficient alternative to the evaluation of actual work samples of past participants' performances would be administration of a good parallel test similar to the embedded test tasks and post tests described in the last chapter. That is, these test tasks or items would be abbreviated to require only a small amount of the participants' time. These tests should be constructed to the same specifications as indicated earlier for pre-test, embedded test, and post test items. Because an assemblage of such items could be completed fairly quickly, and because the items could be cast in such a manner that they could be scored objectively by use of a multiple choice format or some similar short answer objective format, the course developers could very quickly evaluate participants' performances. What would result is some indication of how much retention and growth occurred in the use of course concepts and skills after return to and use of the course content on the job.

It is important to note that one would want to insure that those persons sampled for any type of delayed post testing procedure are actually engaged in work activity which calls for or requires the application of the course concepts and skills which were instructed in the short course. If one sampled many persons who worked in areas which did not involve the use of the course content in their work, one would expect no additional growth in skill or knowledge level. Therefore, any delayed post testing, or other type of assessment of the stability and growth of course concepts and skills, should always be accompanied by procedures for obtaining some other types of information about the individual's use of the course content in his or her work.

Questionnaires, Surveys, and Interviews as Delayed Assessment Procedures

Questionnaires asking about the frequency of use of course content directed to the past participants or their supervisors are very appropriate. So are other questions asking about the critical nature of the course content and skills. Examples of such questions include, "How frequently do you use the content, skills, and procedures learned in the Hydrology and Sedimentology course in your work?" or "In the last three months, how many times have you been faced with a problem where some knowledge, skill, or procedure encountered in the Hydrology and Sedimentology course was essential to the solution of the problem? Appendix A contains many examples of these types of questions which can be used to collect information from past participants and their employers concerning the degree to which the course content is meaningful and centrally involved in ongoing work performance. Past course participants should be asked questions about the degree of critical usage of the course content and skills, the frequency of use, the degree to which participants have recommended other colleagues' attendance at replications of the particular short course, and how useful the course was to their ongoing work acticity. The responses to these types of questions are very informative. Information of this type should be routinely gathered and used along with test data as part of the delayed assessment of the effectiveness of short courses and other formats of continuing engineering education courses.

Data on these types of dimensions can be collected from systematic interviews conducted with samples of past participants or employers by telephone. Written questionnaires and surveys may also be used. In either case, engineers and their employers are usually willing to participate in such activities if it is clear to them that the persons conducting the survey are seeking information about the value of particular short courses.to better meet the needs of practicing engineers. For the same reason, these groups are willing to participate in the completion of formal post tests on course content and skills and to submit actual products of on-the-job work performance for evaluation by course instructors.

Role of Delayed Assessment Activity in "Needs" Assessment

Systematic gathering of such information makes the tacit evaluation of a given short course very clear to the developers of the course and to those who direct and operate continuing education programs. In addition, routinely seeking such information from samples of past continuing education course participants and their employers conveys to both groups a sincere interest in the needs of practicing engineers on the part of continuing engineering educators. Needs assessment is an often used and abused term in the jargon of continuing education. It is often implied that all that is necessary is to "go out into the field" and find out what it is that prospective clients for continuing education courses "need" or "want". There is no better way to be involved in needs assessment activities other than continual interaction with past participants and their employers concerning the outcomes, intended and unintended, of courses already in operation. If courses are developed and operated which serve well the functional needs of professional engineers, and if follow-up activities on the part of the persons who operate continuing education programs convey a sincere interest in making these courses even more effective, new opportunities to develop additional courses and continuing education experiences will arise. The value of these follow-up interactions with participants and their employers is high for both the summative evaluation of the degree to which given courses have achieved intended long term goals as well as for maintaining an open and easy communication between the staff of continuing education programs and the clients they serve.

Conclusion

Oftentimes persons in professional licensing organizations and academic circles tend to attribute more credibility to the results of formal test scores given at the end of courses as measures of learning outcomes than to information gained from the types of follow-up and delayed assessment procedures which have been described in this chapter. In reality, both types of information are needed if one is to make reasonable inferences about the degrees of learning which results from a particular short course or similar continuing education experience.

Chapter 10

DEVELOPING VALID AND RELIABLE TESTS FOR ASSESSING LEARNING OUTCOMES

Earlier chapters have set forth many procedures for making valid inferences about learning outcomes resulting from continuing education courses. There are other ways of measuring learning outcomes resulting from continuing engineering education courses besides formal testing. Some of these alternatives have been described previously. Yet, formal testing is one way by which valid and reliable assessments of learning can be made in an efficient manner. Developing tests which perform in this manner is a demanding task requiring much time and expertise. However, if a particular short course or other type of continuing education instructional activity is to be replicated many times, it is worthwhile to develop pre-tests, embedded tests, post tests, and delayed post tests. The utility of these tests is primarily for assisting in the business of instruction by determining what it is participants have or have not learned and which methods and procedures are most effective in achieving desired learning outcomes.

The remainder of this chapter presents a set of procedures by which to construct good tests useful for making inferences about levels of skill and knowledge in specific performance areas. The procedures apply to the development of pre-tests, embedded tests, post tests and delayed post tests. If the procedures are properly carried out, a large pool of good test items can be developed. These test items can be assembled into parallel forms of the same test. The parallel forms should be the same in terms of the performance capabilities that they test for. They should be different in that the individual items and tasks of which they are comprised, although drawn from the same performance domains, represent different problem situations. If test items are properly developed, it is possible to use the parallel forms of a test interchangably for certain pre-test, post test, and delayed post test functions. This means that the effort expended in developing a good test item pool for a frequently taught course is a good investment although the initial development of the item pool is costly.

Subsequent sections of this chapter deal with each of the first four main steps in a procedure for developing valid and reliable tests. Chapter 11 deals with the fifth step. All five main steps and the various substeps are presented in Table 3. The procedure listed is based on many years of test development activities by many persons. There is a rich theoretical and empirical literature which supports the procedure outlined. The presentation in these two chapters is simplified and basic. Persons wishing more detailed and technical presentations may wish to refer to other sources such as back issues of the Journal of

Table 3

Steps for Developing Pre-Tests, Embedded Tests, Post Tests, and Delayed Post Tests, by Which to Estimate Learning Outcomes

I. Stating Course Objectives in Performance Terms

Determine desired course objectives.

a) List the specific desired learning outcomes in terms of specific performance capabilities.

b) State operational criteria by which adequacy of the performance is to be judged.

II. Mapping Test Items to the Full Range of Performance Objectives

Identify, collect, design, and sample from realistic problem areas test tasks by which to measure the specific performance capabilities listed in step one, across all topics and skill areas.

a) Construct a performance by test item matrix to insure proper and complete coverage of performance objectives by test item tasks.

b) Examine the test tasks which have been selected to insure they are parallel to instructional tasks by which the performance capabilities are to be instructed.

c) Examine the test tasks which have been assembled to insure they have been properly abbreviated from more complex real life problems. Test items must test for critical knowledge and skills in the performance of "real world" tasks. Yet, they must be able to be completed in a short time, usually a few minutes at most.

III. Externally Validating Test Items and Tests

Validate the test tasks initially assembled into a test to insure the test measures what is being taught in the course.

a) Identify other persons expert in the content of the course.

b) Have each expert examine the course content, objectives, and instructional tasks against the array of test items to locate any areas of omission or non-compatibility between instructional tasks and goals and test tasks by which to measure learning toward these goals.

Table 3 (continued)

c) Identify another small sample of persons expert in the content and performance capabilities of the course.

d) Administer the initial assemblage of test items or tasks to this second group of experts. Score their performance on each item. Note problem items where the experts answer incorrectly or have difficulty.

e) Interview each expert and secure his or her suggestions for the improvement of individual test items and the overall collection of items on the test. Revise items and the test accordingly.

f) Administer the test items to a naive group of persons who have an engineering or scientific background but no particular expertise in the area of the performance outcomes taught in the course under consideration. Score the performance of each naive person. Note items which are answered correctly by this naive group. Insure that most of the items on the test are not of this basic or prerequisite type. Rewrite or prepare new items that test for knowledge and skill taught in the course, rather than some more general knowledge or performance capability.

IV. Assembling Items Into Different Types of Tests

Examine each test item which has been developed and sort it into one of these categories, from the standpoint of the level of knowledge or skill required for successful completion of the item. These categories include:

1) Items which test for only basic knowledge and skill assummed upon entry into the course, required for success in course learning activities, and therefore, not instructed in the course.

2) Items which, in addition to 1, test for knowledge and skill outcomes which are the specific performance objectives for the course and which all persons should be able to respond to correctly after completion of the course.

3) More difficult items which, in addition to 1 and 2, test for knowledge and skill that may be expected to improve with practice after the short course is completed because of application of course principles and knowledge on the job.

Table 3 (continued)

a) Sort all items into one of the three categories.
b) Determine the number, variety, and function of all tests needed for assessment procedures including pre-tests, embedded tests, post tests, and delayed post tests.
c) If desired, assemble items from category one into a pre-test. This test would serve only as a screening and advising device. It would be administered prior to a participant entering a course and its results used to make judgment about sufficiency of preparation for course activities.
d) Assemble a comprehensive pre-test, a comprehensive post test, and a comprehensive delayed post test by using items from all three categories. For each test most of the items should be from category two, but there should be some items from category one, and 3 as well to insure a proper range of easy and difficult items. Insure that each of the three tests, with respect to each other, has an equal number of parallel items from categories 1, 2, and 3.
e) Administer and use the three assembled comprehensive test forms interchangeably as pre-tests, post tests, and delayed post tests for different groups of participants on replications of a course, or use each test form for one-third of the participants in any given large enrollment course -- the pre-, post, or delayed post test role.
f) Assemble any embedded tests useful for ongoing assessment of learning during instruction by drawing items from categories 1 and 2. Note that in later stages of learning, items which were initially in category 2 move to category one. Be sure to use items parallel to but not identical to the items sampled for the three forms of the comprehensive tests to avoid "teaching for the test."

V. Conducting Item Analysis and Test Reliability Studies

Using the data collected from test administrations to actual groups of course participants:

a) Determine item difficulty to each item.
b) Determine the ability of each item to discriminate between persons with a good understanding of the subject and those with a poor understanding of the course content.

Table 3 (continued)

c) Determine the reliability of the tests which have been assembled by various procedures, modified to be appropriate to criterion referenced testing and mastery learning approaches if that is the intent of instruction.

d) If needed, modify and rewrite individual test items to produce appropriate levels of difficulty and discrimination across items.

e) Compare the various forms of tests which have been developed, such as the comprehensive pre-test, the comprehensive post test, and the comprehensive delayed post test. If the tests are parallel but consist of different items, each may be considered an independent measure of the same performance capabilities to the degree that all three test forms are highly correlated with one another in any one role. That is, if all three forms are used with three different groups as a comprehensive pre-test, the same results should be obtained by all three forms in this pre-test role. There should be no significant difference in pre-test scores of equivalent randomly selected groups of beginning course enrollees on any of the three test forms. Likewise, any of the three test forms which have been developed should achieve the same results with course enrollees when alternate forms are used in any one role (post test or delayed post test) over replications of course offerings.

f) Plot individual student and group mean scores for pre-test, post test, and delayed post tests against rank order of participants on some external criterion of performance, or against rank order of the post test, across replications of the course with different groups of participants. If the course and the tests are designed properly, post tests and delayed post test scores should be higher than pre-test scores on comprehensive pre-tests. If this pattern does not occur there is a serious problem with the test or with the instruction or both. How high the performance scores should be to demonstrate mastery on the test forms is a matter of logical determination. Mastery may be set to be equivalent to the average score obtained by experts in the course content in step III above. Mastery may also be set at an arbitrary level such as an average of 85 percent correct completion of all test tasks.

Educational Measurement, the leading journal in this area. Many good books are available such as Thorndike's (1971), *Measurement and Evaluation in Psychology and Education*; Maratuza's (1977), *Applying NOrm-Referended and Criterion Referended Measurement in Education*; Nunnally's (1972), *Educational Measurement and Evaluation*; Marshall and Hales' (1972), *Classroom Test Construction*; or Tyler and Wolf's (1974), *Crucial Issues in Testing* to name only a few. The American Educational Research Association Monograph Series on Curriculum Evaluation, volumes 1 through 5 published by Rand-McNally, is another source of excellent articles on this topic. (AERA Monograph Series, 1967-1970).

The procedures presented are ideal. In actual practice, one cannot always employ all of the steps listed. However, attention to the procedures will help insure better tests by which to measure learning outcomes of any course.

Stating Course Objectives in Performance Terms

The first task in the development of a learning assessment test for any course, as well as the development of the course itself, is to specify the specific performance capabilities which are to result from instruction. The particular performance capabilities of the student following instruction should be listed in "action verbs" (Gagne, 1965; Gagne, 1967; Gagne & Briggs, 1974; Manning, 1970; Webb, 1970). Action verbs clearly point out what it is the person will be able to *do* afteriinstruction. They provide a convenient way to capture the essential features of a performance in operational terms. Any one action verb is not sufficient to describe a complex performance in operational terms. However, a carefully selected array of action verbs can often provide a good guide to the selection of instructional activities by which to *teach* the essential elements of the performance and also the selection of test tasks by which to *assess* the degree to which the performance has been learned. Examples of action verbs include *calculate*, *design*, *construct*, *compare*, *recall*, *select*, *organize*, *recognize*, and so forth. How one uses these or similar action verbs to describe the specific performance expected to result from the instructional activity is illustrated in Table 4. The action verbs listed are those for the first unit in the "Hydrology and Sedimentology" course (Hahn & Barfield, 1978) described in earlier chapters. The objectives listed in Table 4 are the basis for the sample test items in Table 5.

Each objective in Table 4 includes one or more action verb in a statement of what the learner should be able to do given certain problem conditions and available resources in the forms of formulas, computational algorithms, charts, and tables. Collectively, the objectives define the expected performance outcome capabilities which should be achieved by students at the end of the unit of instruction. Furthermore, these descriptions are stated in very operational terms. The test items in the sample test in Table 5 and Appendix B are

Table 4

Performance Objectives for Open Channel Hydraulic Structures Unit: An Illustration of Test Construction Procedures*

Objective Number	Action Verb(s)	Description of the Performance Required and the Conditions Under Which it is to Occur
1	Describe	What happens to the value of Manning's n when the boundary of a channel varies through a range of structural conditions including different types of vegetation, non-vegetated soil aggregates, and man-made lining materials.
2	Recall, Recognize	The typical profile of flow velocities (fps) for hydrologic channels of various cross section shapes at typical slopes.
3	Describe Adjust Calculate	The relationship between retardance and flow rate in an hydrologic channel and make adjustments in design specifications (depth, top width, hydraulic radius, slope, and cross section) to produce desired freeboard and channel performance given changes in retardance or flow rates.
4	Calculate	By the limiting velocity method the permissible flow rate for channels given various slopes, required capacities, boundary conditions, soil types, and channel cross sections.
5	Calculate	By appropriate methods and proper use of tables and charts provided, the value of Manning's n for any type of channel given the boundary characteristics.

---

*See Appendix B for details about how the performance descriptions were developed and how test items were designed to measure each objective.

Table 4 (continued)

Objective Number	Action Verb(s)	Description of the Performance Required and the Conditions Under Which it is to Occur
6	Calculate	The hydraulic radius of channels of differing cross sections according to the appropriate modification of the basic computational algorithms.
[illegible]	Calculate	The design specifications for any given channel including the values Vp, R, S, D, T, and necessary free-board given the specifications for any two of these values and information about soil type, topography, etc.
8	Design, Diagram, Label	A hydrologic channel designed to perform to stated specifications under stated problem conditions similar to those listed in item g above.
9	Recognize	The reasonableness of design specifications obtained as the solution to a particular design problem involving a hydrologic channel given the problem variables.
10	Use Select Doublecheck	Appropriately, computational short cut procedures, computational algorithms, and graphic solutions to complex equations given a variety of problems involving the design of hydrologic channels under widely differing conditions of rainfall, soil type, slope, etc.

designed specifically to measure the degree to which these learning outcomes have been achieved. Study of the sample test items in Appendix B, the accompanying textual material in the Appendix, and Tables 4 and 5 in this chapter provides one illustration about how the essential features of a complex performance may be operationalized and translated into specific learning assessment tasks; in this case, particular test items designed to measure each action verb performance statement.

Mapping Test Items to the Full Range of Performance Objectives

It is very important to the development of a good test to be sure that each of the many performance outcomes expected to result from instruction is tested for by several different items. This required an array of items which have been fitted to or "mapped out" to the full range of specific performance objectives. This process of mapping out items to cover all the main performance objectives at their differing levels of difficulty can be aided by a number of systematic approaches. Perhaps an example will be informative.

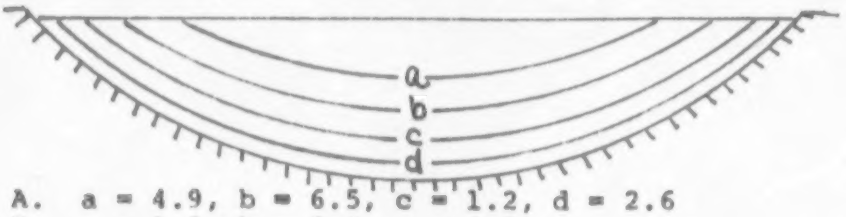
The sample test in Appendix B has eleven items. For convenience the 11 test items are presented in Table 5. The full test with its attached reference materials and detailed narrative explaining how the test items were developed is found in Appendix B. The purpose of the material in the Appendix is to assist the reader in generalizing the

Table 5

TEST FOR "OPEN CHANNEL HYDRAULICS" UNIT - Illustrating the Mapping of Items to Performance Objectives

---

1. What is a typical profile of flow velocities (fps) for the channel cross section represented in this figure?



A. a = 4.9, b = 6.5, c = 1.2, d = 2.6
B. a = 1.2, b = 2.6, c = 4.0, d = 6.5
C. a = 6.5, b = 4.9, c = 2.6, d = 1.2
C. a = 6.5, b = 6.2, c = 2.6, d = 2.3

2. What happens to the value of Manning's n when an erodible parabolic cross section open channel is vegetated compared to an identical nonvegetated channel?

A. increases
B. decreases
C. remains unchanged
D. varies with runoff volume

3. A nonvegetated trapezoidal channel through a sandy loam collidal soil has originally been designed to carry 8 cfs of water down a 4% slope. Suppose the engineer later decides to use a vegetated channel. What must he do to insure an equivalent capacity with the vegetated channel given the same slope, soil conditions, and channel shape?

A. Select a grass which will grow to a uniform height without clumping to assure uniform flow rates at the channel perimeter.
B. Design a somewhat deeper and wider channel to allow for the increased retardance of the flow caused by the vegetation.
C. Design a somewhat shallower and narrower channel because with vegetation a higher flow rate can be sustained.
D. Maintain the original specifications for the non-vegetated channel because the flow capacity will remain nearly unchanged.

Table 5 (continued)

* A channel is to be designed to carry 11.6 cfs of clear water down a 7% slope. The channel material is shale and hardpan. The channel is to be trapezoidal with a 3:1 side slope. Use this information to answer questions 4-8.

4. Using the limiting velocity method, what is the permissible velocity (fps) for water flowing in this nonvegetated channel?

   A. 6.0
   B. 3.5
   C. 2.7
   D. 4.0

5. What is the value of Manning's n for this nonvegetated channel?

   A. .037
   B. .020
   C. .030
   D. .025

6. Using Mannings equation, $Vp = \frac{1.49}{n} R^{2/3} S^{1/2}$, the hydraulic radius of the channel is calculated to be 1.32 ft. The channel cross section area is found from $A = Q/V$ and is calculated to be 1.93 ft^2. The engineer then assumes that the channel depth should be approximately 1.3 feet. He also assumes that the bottom width, d, can be estimated from $A = bd$ where $b + 1.93/1.3$ or 1.48 ft. What should he do next?

   A. Add 20% to the depth value and the bottom width value to provide adequate freeboard in case of a heavy rainstorm.
   B. Check to see if his approximations for depth and bottom width are reasonable by using the relationship

   $$R = \frac{bd + zd^2}{b + 2d\sqrt{z^2 + 1}}$$

   C. Calculate the top width of the channel by using the relationship, $t = b + 2dz$.
   D. Calculate the wetted perimeter value for the channel using the relationship $2d\sqrt{z^2 + 1}$ to determine flow resistance.

---

*Items enclosed in brackets contain information in their stems necessary for the solution of problems contained in later items in that group of items.

Table 5 (continued)

7. What can be said about the engineer's estimates of the values for the depth and bottom width of the channel?

    A. Both values are a reasonable approximation of the true values.
    B. Neither value is a reasonable approximation of the true value.
    C. The width estimation based on assuming a rectangular cross section is only slightly in error
    D. The depth approximation is based upon assuming that R = d and is quite accurate for this channel.

8. What are the final values which are necessary for the depth, (D) bottom width (b), and top width (T) of the channel if it is to operate at the capacity given in the first part of this problem and under the soil and slope conditions specified? Include the necessary freeboard (ft.).

    A. D = 1.3, b = 1.5, T = 9.26
    B. D = 1.6, b = 1.8, T = 11.1
    C. D = 2.0, b = 7.0, T = 15.0
    D. D = 2.4, b = 7.0, T = 18.0

A parabolic channel is to be designed to carry 25 cfs of water on a 4% slope. Because the soil is easily eroded, the designer decides to vegetate the channel with fescue which is to be unmowed. Use this information to answer questions 9 - 11.

9. What is the maximum permissible velocity for water flowing through this channel (fps)?

    A. 3
    B. 5
    C. 7
    D. 3.5

10. What is the retardance class for this vegetated channel?

    A. A
    B. B
    C. C
    D. D

11. What is the hydraulic radius of this channel?

    A. 1.1
    B. .58
    C. .82
    D. 1.6

principles demonstrated in this illustration to the design of other tests. The test items in Table 5 were written to test for the presence of the performance objectives listed in Table 4. Examination of the test items in the sample test and the performance objectives for which they were generated reveals several important points.

First, it is apparent that the objectives and the test items written to measure their achievement vary in difficulty and level of complexity. The first items and objectives are simpler, the latter more difficult and complex. The first objectives and items mainly call for description, recall, and recognition of certain relationships, facts, and principles. Later objectives and items call for application of knowledge of facts, principles, and relationships in the actual solving of some complex problem.

A good test should always consist of an array of items across levels of performance from complex to difficult. Such a test reveals much about what parts of the final performance have been learned well or not so well. Learning complex performance tasks is a gradual process. It takes much time and practice. It can be expected that early assessments with test tasks similar to those displayed in the sample test in Table 5 will show stronger performance in knowledge of facts and recall of particular formulas than in other areas dealing with the skillful and rapid integration of all the subskills and knowledge into a complex problem solving activity. Thus, as student performance is examined across pre-tests, embedded tests, and post tests, there should be much growth, particularly in the items dealing with skillful application of course content and skills. If such growth does not appear, there is something wrong with the instruction or with the test.

One common mistake made with test items, especially of the multiple choice variety, is to include only low level, recall of factual information types of items. A test composed of only these types of items is very limited as an assessment of learning outcomes of a more complex nature. Inferences about the overall degree of learning resulting from instruction based on tests are valid only so long as the test items accurately represent the full range of objectives from simple to complex. Each performance outcome intended as a goal of instruction must be represented in the test items presented in order to allow the student to demonstrate what he or she has learned.

Inspection of the sample performance objectives, listed in Table 4 and the sample test (Table 5) for those performance objectives, reveals another interesting characteristic. For the particular objectives written for the unit, the test items presented are only one possible sample of the many items which could be written to test for students' skill in the achievement of the intended performance outcomes. For any given performance objective which is stated in operational terms, it is possible to design many different test items which measure, to some degree, the underlying performance of interest. This is why it is relatively easy to develop many similar items for use on parallel test forms for use as pretests, embedded tests, post tests, or delayed post tests. This property of multiple test items for each performance objective becomes more pronounced as the performance objectives become more broad and inclusive. Thus, it is even easier to write many parallel form test items for complex and high level performance objectives than for simple performance objectives which require only the memorization and recall of specific information.

Inspection of the sample objectives and the test items developed for these objectives also reveals another important relationship. Any given test item may test for more than one specific performance objective. Complex and demanding test items usually test not only for the ability to integrate and apply much knowledge, skill, and judgment into the correct approach and solution of a problem. In addition, they test for knowledge and recall of appropriate facts, relationships, procedures, and skills in bringing all of these components together appropriately into a competent and skilled performance. Items 4 through 8 and 9 through 11 in the test in Table 5 illustrate this property. Each of these items tests for knowledge at multiple levels. It is desirable to develop and use such multiple performance assessment test items. However, for every test item it should be clear what particular performance outcomes are being measured. Care must also be taken to insure that the performance demanded by the item can be completed in a short and reasonable length of time for one test item. An earlier chapter has provided guidelines and examples of how to abbreviate such complex performance assessment tasks without losing the essence of the main features required for a skillful performance. The additional explanatory material in Appendix B for the test in Table 5 also provides information about how to abbreviate complex performance tasks to make them into test items.

Taxonomic and Task Analysis Approaches to Mapping Items and Objectives

There are many procedures for monitoring the appropriate level and sequencing of a hierarchy of performance objectives and test tasks in the design of instructional activities and the development of a range of test items by which to assess achievement of the objectives. One good general guide to striking an appropriate balance of objectives and items, from a simple and fact oriented to a complex and skill oriented focus, is the Taxonomy of Educational Objectives: Handbook I, Cognitive Domain (Bloom, 1956). This manual provides many specific suggestions and illustrations of how to assemble reasonable samples of both performance objectives and test tasks and items by which to assess performance across various levels. Another source which provides detailed procedures and examples of how to construct arrays of performance objectives and test items is the Handbook on Formative and Summative Evaluation of Student Learning (Bloom et al., 1971).

Another approach to the problem may be found in task analysis interpretations of complex performances. Under this approach, the final performance capability intended as the outcome of instruction is described in all of its complexity. Then one asks oneself and others expert in the performance domain, "What knowledge and skill are prerequisite to being able to produce the final skilled performance?" In this manner, a series of knowledges and skills are conceptualized in a descending hierarchy from very highly skilled and complex performance to very basic levels of knowledge and skill. After the conceptualization of the complex performance is "task analyzed" the various sub-skills and knowledges become the focus of particular learning objectives which are now arranged in an ascending hierarchy. The intended performance objectives at each level of the hierarchy are used to define the content, topics, learning activities, and instructional materials which will be used to instruct participants up the hierarchy of knowledge and skill. The same conceptual hierarchy is used to develop a series of test or assessment tasks which parallel the instructional tasks. The task analysis approach to the design of instructional objectives, instructional tasks and activities by which to teach the performance objectives, and the test items by which to assess learning resulting from instruction, is particularly well suited to technical and scientific fields. This is because in these fields there is a natural cumulation of information, concepts, skills, and complex procedures into integrated and very high level performances (Gagne, 1962; Gagne, 1967; Gagne & Paradise, 1961; Salvendy & Seymour, 1973). The sample performance objectives in Table 4 and the sample test items in Table 5 were developed in a task analysis manner and provide a concrete illustration of the procedure.

Persons interested in the task analysis approach to the design of instructional objectives, instructional activities, and assessment tasks will find The Conditions of Learning (Gagne, 1977) and Principles of Instructional Design (Gagne & Briggs, 1974) to be of value. Both of these texts provide many details for using task analysis procedures. In addition, many studies which are excellent illustrations of how to use the procedure in the design of technical courses are referenced. Many of the examples are from the technical and scientific fields. Another good source is a dissertation titled, An Evaluation Model for Developmental Growth (Marion, 1978). Marion's research is particularly interesting since it deals explicitly with the measurement of complex performance capabilities resulting from technical training programs. In addition, Marion's procedures are explicitly conceptualized with a learning hierarchy framework and designed to integrate diverse types of information from areas such as formal test scores, actual student performance in laboratory or clinical settings, and instructor or supervisor ratings into a common assessment of the degree of student learning.

Either the taxonomic approach of Bloom or the task analysis approach of Gagne, Marion and others, can produce a set of performance objectives and assessment tasks which are very operational. These approaches can also be nicely integrated with one another. Either approach also tends to produce a proper array of test items representative of the full knowledge and skill domain required for adequate performance of complex tasks. This is especially so if attempts are made to ground performance objectives, instructional activities and tasks, and test tasks in real world problem solving tasks typically encountered in practice by the engineer (Salvendy & Seymour, 1973).

Constructing a Matrix of Objectives by Test Items by Topics

No matter what approach is used to generate test items, it is a good idea to construct a matrix of objectives by items for each topic included in a course. This insures the development of better criterion by which to measure learning outcomes (Manning, 1970). The particular performance objectives may be listed in a column, similar to the presentation of the sample objectives for the unit in the

Hydrology and Sedimentology course in Table 4. Test items which have been developed can be sorted into rows corresponding to each level of objective represented in the column. This procedure quickly reveals any imbalance of items in the total array to be used for the test. If too many items are written at the information level and too few at the application of principles level it becomes immediately apparent. Gaps often will be identified for levels of objectives for which there are no items, but for which items need to be developed if the test is to be representative of the range of content and skills instructed. It is also possible to continue this same matrix construction for other content areas of a given course. For example, in the "Hydrology and Sedimentology" course, one could use the same hierarchy of action verbs in a sequence of objectives across, not only the first unit in the course on open channel hydraulics, but for each of the other five units in the course as well. A separate matrix may be constructed for each major topic or unit in the course. What results is a matrix of objectives by test items by topics across the course. In this way, the matrix of specific performance objectives at various levels by content or topic areas can be completely laid out. With such a structure it is much easier to map out a representative set of test items capable of providing a good assessment of the degree of learning achieved by students following instruction. Persons interested insuring an appropriate coverage of all performance and content areas by test items, there remains still another task before the test is assembled. It is necessary to insure that each test item can be completed in a short period of time. As was indicated in Chapter 7, test items and tasks must be abbreviated situations which call for the performance capabilities required in applying course content and skill to the solution of real problems in some complex performance domain. The test must always be much shorter than the time available for instruction and the time available for formal instruction must always be shorter than the time available in the work setting for the actual use of concepts and skills which are the focus of instruction. Unless one is careful to abbreviate test tasks and items, one ends up with assessments of learning which are very incomplete since all of the available testing time is spent on the completion of one or two test tasks. If the test tasks are complex and require a comprehensive integration of course knowledge and skills, the test can be more representative of the performance domain being taught. However, it is generally better to have multiple and independent indicators of learning outcomes by which to make inferences about persons' competence in complex performance areas. Thus many shortened or abbreviated test items are usually preferable to only one or two major and time consuming test tasks. One exception to this rule is the laboratory practical examination where performance is assessed by directly observing the person design and conduct an experiment, construct a computer program, or analyze the chemical properties of an unknown material. However, for many courses multiple short item tests remain the most practical and useful means of estimation of learning achievement short of actual observation of on-the-job performance and the products resulting from this performance.

Chapter 7 outlined many procedures by which complex performance tasks can be abbreviated into efficient and brief test tasks or items. Consequently, no more will be said about this matter here.

Externally Validating Test Items and Tests

All of the previous activities are directed toward developing tests which are valid indicators of the learning outcomes which result from a course. However, so far, all of the validation procedures are internal. They are based upon the judgment of the course developer and test item constructor. In actual practice the course developer and test item constructor tend to be the same person or persons. This is perfectly normal and desirable. As is pointed out in Chapter 7, test items and instructional tasks are closely related. Both are drawn from the same domain of performance capabilities. Test items should be rooted in the same array of content, knowledge, and skill as are the instructional tasks which are selected for teaching a course. However, before a test is developed for repeated and wide use with many replications of a course, the validity of the test for the performance domain being dealt with should be checked by persons external to the course and its development.

There are two basic ways to approach this external validity check of initial forms of tests which have been developed. The first way involves identification of a few persons who are very expert in the content and skills of the course. The number can be quite small, consisting of only two or three persons. These persons can be given the performance objective by item by topic matrix as well as the actual assembled test which has been constructed initially. In addition those experts should also be given a description of the teaching methodology and a set of the course instructional materials. These experts can then be asked to examine all of these materials and make observations about the adequacy and scope of test items (and instructional tasks and objectives for that matter) in terms of their own expert opinion of what is required to exhibit skilled performance in the content area under consideration. The expertise of these external reviewers allows them to make reasonable and independent judgments about the adequacy of test items by which to make inferences about a person's ability in the content area under consideration.

There is another important reason to interview the experts who have studied or actually taken the test. If the test items are too easy, and if they represent only the basic and entry level skills required for skilled practice in some complex performance area, test score data from the group of experts would by itself not reveal this weakness. Yet a dialogue with the persons completing the test would almost certainly do so.

Another strategy is to locate a group of persons naive in the particular performance area for which the test has been developed. These persons should have some common technical background and training compared to the persons who are expert in the area. However, they should lack the particular training, experience, and background in the performance area which is the focus of the short course for which the test is developed. This naive group can be administered the initial test which has been developed for the course. Obviously, the performance of these persons should be poor in terms of mastery of the test material. If this group of naive persons performs at high levels on the test, it means that the test items are not a valid indication of what has been learned in the course. There may be problems with too many low level items being included in the test or with other high level items being constructed in such a manner as to reveal the correct answer or the way to obtain the correct answer. In such an event, the test items need to be reworked. Once again, interviewing of individuals who obtained high scores on the test, without apparent prior knowledge of the content area, can provide much information about how to revise existing items or construct new items to be more valid measures of particular performance objectives.

It should be emphasized that the numbers of persons in these groups of expert judges, expert test takers, and naive test takers need not be large. If one has only two expert judges independent of the course development activity, only five to eight expert test takers, and only five to eight naive test takers, much can be learned from the results which will improve the validity of the test being developed. It is far better to use such small and independent samples by which to externally validate a test which is under development, especially if one interviews the persons in these groups after their activity, than to use large numbers of persons and rework test items only on the basis of the test scores and individual item characteristics. Groups of these compositions and sizes are very adequate to the task of identifying the more serious problems with test items. Usually any serious problems will be identified by multiple persons in the expert group even with only a few persons involved.

Assembling Items Into Different Types of Tests

The initial development of the levels of performance objectives by course topics by test items matrix will provide much information about the general difficulty level of the tests. However, another sorting of the pool of test items into three categories can help in assembling tests likely to be effective in discriminating among different levels of achievement of learning outcomes.

The three levels are found in Table 3 on page 161. In the first level are those items which represent knowledge and skill prerequisite to the course and its successful engagement by the learner. The knowledge and skill represented in these test items is believed to be necessary to learning the content of the course. However, it is not to be taught in the course because there is limited available time, much content, and students must be assumed to have an entry level of knowledge and skill in order to proceed.

In actual practice, persons who enroll in continuing engineering education courses vary a great deal in the degree to which they have mastered prerequisite knowledge and skill (Weisehugel, 1978). Therefore, in the interest of documentation of the growth of individuals' learning, as well as the average effectiveness of a course in improving the learning of groups of individuals, it is necessary to have some idea of where participants actually entered in respect to expected prerequisite levels of knowledge and skill. For example, suppose a course had been developed for teaching practicing civil engineers how to use computer simulations to test the performance of various structures and systems they normally design. Also suppose that the course was well designed and had the potential for being very effective in teaching practicing engineers methods of computer simulation by which to test the adequacy of complex designs. Let us also suppose that the prerequisite skills require participants to be facile with Fortran or another computer language. In addition, suppose the participants are also expected to be very familiar with the use of computers and computer terminals. Now, suppose 40 percent of the participants for a given offering of the course did not possess these prerequisites to a high degree. These students would undoubtedly have difficulty in doing the course learning activities and would also be likely to perform poorly on any post test which was reasonably representative of course knowledge and skill objectives. If the success of the course were judged on the post test scores alone, it might be seen as a poor course, or at least as not being effective for a large number of persons, which, indeed, it was not. However, without some measure of the entry level skill and knowledge capability of the participants, one could not be sure of the reason for the lack of success. The same result could be achieved from poor organization of instruction, poor presentation of course material, a hostile attitude of the instructor, or distracting and inadequate conditions in the physical setting of the learning environment.

For this reason, it is generally wise to incorporate some items which measure prerequisite skills and knowledge levels in tests developed to assess learning outcomes of courses. This can often be accomplished by a relatively few number of items, perhaps as few as four or five, each item carefully selected to require the performance of some specific prerequisite skill or the recall of some specific procedure or information.

There are two ways to approach the inclusion of items which measure basic knowledge and skills required for entry into and completion of course activities. The first method is to produce a test designed specifically to measure prerequisite knowledge and skill. This type of test can be mailed out to participants in advance or administered in one central place ahead of time. The test can be scored and the results used in an advisory way to place a participant in a course appropriate to his or her needs and present level of capability. The purpose of the test is to screen and advise persons concerning entry into a given course. Persons scoring low on such a prerequisite skills and knowledge test should not be restricted from entering the course in most cases. Rather, they should be advised that they would do well to not enroll in the course at this time, and also be advised about what they needed to do to prepare for the course if they wished to enroll in the future. An exception to this would be in areas where entry into the course without the prerequisite skills and knowledge would be dangerous or destructive to property and life. Until a pilot demonstrates strong proficency on a wide variety of prerequisite knowledge and skill tasks, including written tests, physical performance tasks in a Link trainer or some similar device, he or she should not be permitted to enter the next phase of flight instruction, e.g., actual flying. Similar restrictions operate in the use of complex, expensive, and potentially dangerous equipment as is often found in laboratories and industry.

Another way to provide information on the general entry level skill and knowledge of participants in prerequisites is to prepare comprehensive pre-tests. Under this approach, one prepares a test comprised mainly of items which are sampled from the domains of performance being instructed in the course. Many or most of these test tasks or items would also require prerequisite knowledge and skill. However, if a person missed these items on a test, one would not know if it was because he or she did not have the necessary prerequisites or if the individual had not learned or misunderstood something in the new knowledge or skill area being taught. The way to avoid this problem is to include a small number of items on the comprehensive pre-test which require only the expected prerequisite knowledge and skill for their correct completion, and nothing from the present course being taught. Again, these items can usually be few in number, being the same types of items as would be used in a screening pre-test, but serving a slightly different purpose. Here responses to these items would serve to reveal information about the variability of the entry level skills and knowledge of participants actually admitted to a course. The purpose would be to make better inferences about individual learning and general course effectiveness.

There is another reason for developing and using comprehensive pre-tests. In fact, when such a test has been developed it becomes capable of being used as a pre-test, a post test, or a delayed post test. This is because any comprehensive test includes items from across all three categories shown in Table 3, page 161. Most of the items should be from the second category and be directly related to the intended learning outcomes for the course. These items should be tasks the learner can be expected to perform correctly having completed the course and all of its instructional activities. In addition, there should be a small number of items which test for only key prerequisite skills and knowledge, for the reasons described above. However, there should also be a small number of items from the third category, a sampling of test tasks which the person cannot yet be expected to perform very well, after only having completed the course. These items should permit the learner to demonstrate skill and knowledge beyond that expected to result immediately from the completion of course activities. When such a comprehensive test has been assembled its use in the role of a pre-, post, or delayed post test reveals information about the range of competencies of participants, which are often very variable, especially upon entry to the course.

As in the case of the inclusion of items related only to prerequisite knowledge and skill, only a few items which go beyond the levels of skill and knowledge expected to be achieved in the course need to be included. The presence of these items on the test allows those persons who have high levels of capability to not be restricted by the test. Inclusion of these more difficult items also makes the test useful as a delayed post test. It has been noted earlier that for most continuing education courses in engineering, one should expect the persons who have completed the course to learn even more after the short course is completed and persons have returned to the work setting to apply course content. To be sensitive to this additional facility with the knowledge and skills of the course and their appropriate application, some items of the more demanding type need to be included. It is also usually the case that some participants will have entered with higher levels of knowledge and skill in a particular course than have other students. It is not uncommon for some students to learn more than might be expected from only what is actually instructed.

In summary, a comprehensive test consisting of a sample of items from all three levels provides more information about the entry and exit levels of participants' skills and knowledge on an individual basis and as a means of making inferences about the effectiveness of the course in general. To properly discriminate among persons with differing levels of ability, a test must be composed of an array of items of differing difficulty. Including items from each of the three categories is one way which is likely to lead to an appropriate range of difficulties in the items assembled for a test.

When comprehensive tests are assembled, care should be taken to have an item pool which is several times larger in terms of numbers of items than the length of the tests which are to be constructed. Suppose that one wishes to construct a comprehensive test of about 20 items in length for a particular course. Suppose that it is also decided that a pre-test, post test, and a delayed post test would be useful. In addition, four items are included to test prerequisite knowledge and skill. Four more items, which are very difficult and demand learning beyond the level included in course learning activities, are also included. The remaining 12 items are all related to specific performance objectives taught in the course. If three such parallel tests were to be constructed, it would be necessary to have a minimum of three times these numbers of items in the item pool in each category. Thus there should be 12 items of the prerequisite knowledge and skill type, 36 items related to specific course performance objectives, and 12 items related to transfer of course knowledge and skill to more complex problems not dealt with in the course but which might be reasonably expected to be samples of actual applications following course activities.

One would then construct three parallel test forms from this array of items. First, one of the three parallel items from the prerequisite knowledge and skill item pool for a given prerequisite would be randomly assigned to each of the three forms of the test. This process would continue until each of the triplet of items for each of the remaining three prerequisites was assigned randomly to one of the three forms of the test. The same procedure would be repeated with the 12 triplets of items in the performance outcome item pool. The procedure would be repeated again for the four triplets of items in the transfer pool. What would result would be three parallel forms of one test. Any form of the test could be used in any of the three roles of comprehensive pre-, post, or delayed post test. To the degree that the tests were actually empirically as well as conceptually parallel, a set of three test scores across an individual would determine his or her entry level, exit level, and subsequent improvement or decrement in knowledge and skill in course content at some time after the completion of the course. Over replications of courses and random assignment of persons to test forms the statistical significance of each form of the test can be empirically determined by methods similar to those presented in Chapter 13. If the test forms are found not to be equivalent, adjustments can be made in particular items on various forms to produce parallel items and tests.

It is much more likely that parallel items and tests will be developed if one begins with a listing of specific performance outcomes, as is shown in Table 4, and if a task analysis procedure has been used to generate the performance objectives for the course. For short courses this specific mapping out of the key performance aspects of the tasks to be taught and the use of action verbs to operationalize both the instructional tasks and the test items is a sound approach. Each action verb and performance description provided (See Table 4) can easily produce a number of items identical with respect to the performance being tested, but unique in terms of the specifics of the problem situation. Each of the parallel items thus produced may then be randomly assigned to any test function or form.

The actual degree of which test forms are parallel can be determined by a variety of means, the most common ones being based on analyses of group means and variances for comparable groups of participants for given roles of the test, e.g., pre-test, post test, or delayed post test. There are two easy ways to do this. Over subsequent replications of a course, one can use randomly each form of the test in a different role. Thus for the first group, form A can be used as a pre-test, form B as the post test, and form C for the delayed post test. For the next replication of the course the order can be reversed. Still later, replications can be used to assign form B to the pre- or delayed post test role. If it can be assumed that the groups of participants are comparable, the means and variances of any form of the test ought not to be statistically significantly different from one another when used in any particular role. This should be especially true for the post test role where participants immediately following completion of the course and having had a common developmental or learning experience, are likely to be most alike.

An alternative is to use all three forms of the test in all three roles for each course replication. That is, one third of the participants randomly are assigned form A for the pre-test; another third, form B; and the remaining third, form C. The same pattern is followed for assignment of persons in the course for the post test with the constraint that no individual may be assigned the form he or she was previously assigned. The delayed post test assignments are made in the same manner. Again, if there are no significant statistical differences between tha means of the three forms of the test in each of the three roles, the test forms are supported as being parallel empirically as well as conceptually. Under this second plan, one might have to accumulate data over several replications of a course to have enough persons' scores on each form of the test to make strong inferences about the parallel nature of the tests.

Any embedded test tasks which are to be used ought to be another set of parallel items, not those used for the construction of the pre-, post, or delayed post test. One can economize somewhat by using the pre-test items as embedded test items and reserving the other two forms of the test for the post and delayed post test. The important condition to maintain is to keep post tests independent from the embedded test tasks. The latter are primarily used to teach the student and help the instructor guide practice activities for the student, while the post and delayed post tests are designed as an independent assessment of learned capabilities.

Conclusion

It should be apparent that one does not go through this lengthy process unless it is likely that a given course will be developed and taught often and unless there is strong interest in determining the learning outcomes resulting from the course by estimates using test scores. If the course is only to be developed and taught once or twice, or if there are other good indicators of the functional capabilities of participants after the course is completed, it would make no sense to spend the effort involved in developing comprehensive pre-tests, post tests, and delayed post tests.

Chapter 11

CONDUCTING ITEM ANALYSIS AND TEST RELIABILITY STUDIES

Thus far, all of the test development activities described in earlier chapters have depended upon the knowledge of the course instructors and some other persons expert in the content of the course. If the steps outlined previously are followed, it is likely that tests will be developed which are good estimates of the degree of learning which results from a given course.

However, there are well established test item analysis methods which can be used to calculate the difficulty level of given items and also the ability of items to discriminate among persons who understand the content of the course and those who do not. A difficulty index can be calculated for each item as can a discrimination index. These statistics provide empirical information about the behavior of the items in the tests which have been assembled. This empirical information, coupled with the information derived from the logical development and analysis of items desdribed in earlier chapters, can be used to rewrite and adjust test items to achieve optimal levels of difficulty and discrimination.

In addition, there are well developed methods for the empirical estimation of the reliability of a test, i.e., the degree to which the test consistently measures the presence of given levels of ability in persons from trial to trial.

If tests are not reasonably reliable they are poor estimates of the degree of learning which has occurred. Consequently it is important to determine the reliability of tests and to modify tests to be more reliable if they are not so initially.

Only a brief introduction to some of the more common methods by which to carry out item analysis and reliability studies is presented here. The books referenced later in this chapter provide detailed information about these procedures.

Before proceeding, it is important to realize there is a problem with the computation of item difficulties, discrimination indices, and the reliability of tests by traditional means because of the nature of the tests which have been described in earlier chapters. Most of the standard procedures for carrying out item analysis and test reliability studies have been developed for norm referenced testing approaches. Yet, this book describes and recommends a criterion referenced testing approach for developing tests by which to assess learning outcomes of continuing engineering education. This means that typical item analysis and test reliability estimation procedures, while useful to the task of test construction, must be modified. It also means that persons using these methods need to be clear about the origins and assumptions implicit in them. After examining this problem, suggestions will be made for the use of existing procedures for conducting item analysis and test reliability studies in a manner consistent with the purpose of testing as it occurs in criterion referenced situations.

Norm Referenced Testing Procedures

There are two general approaches to testing. One is called the norm referenced approach and the other the criterion referenced approach (Airasian & Madaus, 1974; Bloom et al., 1972; Maratuza, 1977; Millman, 1974).

The norm referenced approach is based upon obtaining a representative sample of persons from some population of interest. Tests which have been developed are administered to this sample of persons. Judgments about the adequacy of an individual's performance are made by reference to the mean performance of the group and the observed standard deviation for the sample. Individual persons' performance scores on the test are ranked with respect to one another. "Passing" performance is defined in terms of falling within some range of typical performance around or above the mean score for the group. Arbitrary criterion cut off points are used. For example, it is sometimes asserted that any person whose score falls below the 50th percentile will not be admitted to a program. Sometimes the cut-off point is determined in terms of standard deviations so that persons whose scores on the test fall below three-fourths of a standard deviation below the group mean are judged as not knowing enough to have passed the test.

Norm referenced approaches are commonly used in the construction of standardized achievement tests which are used for making decisions about admission of students to academic programs. Group intelligence tests are also norm referenced. Most professional licensing and certification examinations are also norm referenced. Thus, in a given administration of a professional engineering licensure examination, a certain percentage of the persons who take the test may be expected to fail the test because success is defined with respect to achieving a score no lower than so many standard deviations below the mean of the group taking the examination. Because all test administrations to samples of persons result in some variance and because that variance can be used to define a mimimum passing score which falls at some arbitrary point below the mean, no matter how selective the group being tested and how skilled each person is in the knowledge and skill being tested, some proportion of the persons taking the test will fail by definition.

The norm group for a norm referenced test may be only the persons taking a particular test at a particular time for licensure as an engineer. In practice, the norm reference group ought to be much more inclusive. Norms for well developed achievement tests are usually based upon national random samples of persons from the population of interest. Information about an individual's test score can be interpreted with respect to these national and regional norms in terms of ranking the competence of the individual in the test content compared to these other persons.

A much less adequate version of norm referenced testing has also been the most common practice of professors in engineering, scientific, and technical fields. This has often been referred to as "grading on the curve." In this approach the professor constructs an examination based on the course content. The test is then administered. Grades are assigned by placing a certain percentage of the top ranked observed student scores in the A category, the next group of scores in the B category and so forth until a certain percentage of the lowest ranked scores are placed in the F or fail category. Criteria for determining these grade assignments may be based upon simple rank order of scores, percentile ranks, standard deviation units above or below the mean, or other similar procedures. Whatever the procedure there are often serious problems with this norm referenced approach based on the performance of the students in only a particular classroom.

First of all, most groups of students in engineering classes are highly selected with respect to their prior knowledge and skill which is required for entry to the program and successful participation in the class learning activities. This is so especially for the students in advanced courses. It is also true for many students who are professional engineers enrolled in continuing engineering education courses.

Suppose that a group of students in an advanced level engineering course in a technical area is already highly skilled and knowledgable in the prerequisites to the course activities. Suppose that, in addition, these persons are also highly motivated to learn what the course has to offer. Also suppose that the course is well organized and the instructor is quite effectice in his or her teaching. Now, let us suppose that the instructor prepares an examination to test the students' knowledge of course content. Let us assume there are 16 students in this course. We will assume it is a well designed and reliable examination. Let us also assume that most students have worked hard and have, indeed, learned most of the content and skills of the course. Each student completes the 25 item examination. The test consists of some very difficult items, a few easy items, and other items of moderate difficulty for persons of this general ability level. The mean score for the group is 17, and the standard deviation is 2.3. Would it make sense to fail those students whose observed score on the test happened to be ranked last or happended to be one or one and one-half standard deviations below the class mean? What would be the meaning of the grade assigned to each student? Suppose the next semester the same course is taught in the same way by the same instructor to another group of 13 students. The same examination is administered. This time the mean for the group is 13.2 and the standard deviation is 4.4. Once again the instructor assigns grades by rank ordering the students' observed scores in the class. He uses the same criterion of so many standard deviations below the mean as indicating failure. What do the set of grades in the second class mean? In particular what do the sets of grades for persons in the two classes mean with respect to one another? Not very much! In both cases the normative reference group is non-random, non-representative of the larger population of persons at that level of development and expertise, and too small and truncated.

Norm referenced testing procedures make good sense for determining the skill or competence of a person in comparison to other persons in the general population of interest. As such the norm referenced approach makes sense in the development of standardized achievement tests based on national or regional samples of the population of interest, insuring that persons from all ability levels have an equal opportunity to be sampled in the norm group. Without random sampling of persons across ability levels the individual score of a person cannot rank him or her with respect to the distribution of knowledge and skill in the population to which he or she belongs. This property of norm referenced tests makes them very useful for standardized achievement tests and very inadequate as a means of estimating the degree of specific intended learning outcomes for courses. Because of these problems with norm referenced testing approaches an alternative has been developed which is much more appropriate to instructional settings for the estimation of the success of instruction in achieving specified learning outcomes by individual students.

Criterion Referenced Testing Procedures

Particular courses have specific intended learning outcomes which are usually much more precisely and narrowly defined than the domain of knowledge and skill typically tested for on a standardized achievement test. In a particular course, the instructor usually wants to teach some finite number of specific facts, concepts, principles, and procedures as well as skill in actually applying all of these to the solution of problems faced in the real world work situation. This is the case especially for continuing education courses for practicing engineers. There is little interest in comparing the performance of individual students who have completed the course to one another or to some external normative reference group. There is much interest in comparing the performance of students on tasks typical of those faced in the work setting with acceptable standards of safe and informed practice. The student is expected to learn the appropriate use of knowledge and skill acquired in the course in order to exhibit performance of some particular tasks at a criterion of mastery. This approach to testing has come to be known as criterion referenced testing.

In criterion referenced testing each person's performance is compared to some standard of mastery or competence in the actual performance of tasks directly related to the knowledge and skills taught in the course. The criterion for comparison is how well individuals are able to do certain tasks which are specifically sampled to be representative of the range of typical tasks being instructed. These selected tasks are similar to those assigned for practice during the course of instruction. The entire emphasis is upon the degree to which each person, having received instruction, is capable of exhibiting competent performance on specific tasks similar to those used to teach persons in the course but never before encountered in that particular configuration. Testing tasks and instructional tasks are parallel. Both are focused on specific performance capabilities which are seen as the purpose of instruction. There is little or no interest in grading persons with respect to one another, only in estimating how well persons have learned particular complex performances. These performances are usually expressed in action verbs as noted in Chapter 10. They are operational and observable. The acceptable level of "correct" performance is defined as mastery. The mastery level is determined by the actual degree of correct performance which is possible or required at the functional professional level in the real domain of actual work performance. In engineering education the performance-based engineering approach (Grogan, 1979), self-paced teaching methods (Cleaver, 1976), the personalized system of instruction (Kulik & Kulik, 1975) and other similar approaches are mastery learning oriented. All use criterion referenced testing approaches. These methods have been widely used and have been shown to be very effective.

It should be clear to the reader that the suggestions and guidelines which have been presented in previous chapters are directed toward the criterion referenced approach to testing and assessment of persons' functional competencies. The problem is that many of the procedures which have been developed to determine item difficulty, discrimination, and test reliability have been developed under normative testing approaches for the development of very inclusive and non-specific performance tests which test for very broad domains of general achievement.

Item Difficulty and Discrimination Indices for Norm and Criterion Referenced Tests

Under norm referenced approaches to testing a common rule of thumb for determining item difficulties is to divide all of the persons who took the test into a top quarter and a bottom quarter group with respect to observed scores. Sometimes if the total group size is small, the top and bottom thirds of the scores are used. An index of item difficulty is calculated in a straight forward manner

$$\text{NR Item Difficulty Index} = \frac{\text{Proportion of high scorers correct} + \text{Proportion of low scorers correct}}{2}$$

The difficulty of each item is calculated in this manner and the average item difficulty of the test items is determined. The properties of reliability formulas and long experience have shown that tests with average item difficulties of .5 produce the optimal range of scores and result in optimum reliability of the test. Average item difficulties of this magnitude are ideally suited to the task of separating people and ranking them in terms of test scores obtained. In actual practice test items which have difficulty indices ranging from .25 to .75 are desirable for this separation and ranking function. In norm referenced testing it is desirable to end up with a test which has a nice range of item difficulties with the mean item difficulty being about .5. If the item difficulties are too high, the test will not permit students to demonstrate what they know, the scores will all be lumped together at the low end of the scale, and students' test scores will not be suitable for separation and ranking. If the item difficulties are too low the test will not assess what students do now know, the scores will be all lumped together at the high end of the scale, and, again, not be suitable for separation and ranking. But ranking in respect to what? With respect to other persons, of course, for that is the matter of interest in norm referenced approaches to test construction.

In a criterion referenced approach there is little or no interest in comparison or ranking of persons' scores with one another. There is strong interest in ranking or categorizing persons' scores in relation to some criteria or standard of competent performance on some sample of tasks similar to those used in teaching the knowledge and skill. Typically the criterion is arbitrarily established and called a "mastery" level. Mastery is usually defined in terms of a certain percentage of correct performances over trials or tasks. For example, the criterion for a mastery of a course might be absolutely correct performance on at least 80 percent of all test tasks given on any particular test. Less than 100 percent correct performance on at least 80 percent of all test tasks would be viewed as failure to achieve mastery, while 100 percent correct performance on 80 or a greater percentage of all the test tasks administered would be viewed as demonstration of mastery.

In the criterion referenced testing approach it makes little sense to compute item difficulties in the traditional manner. A better procedure is to administer each test task to groups of persons who have not completed the course of instruction and who should be naive in the skills and procedures being taught. If this is so, these persons should perform poorly on the test tasks or items. However, a second group of persons who have completed the course, and/or who, independent of course completion are known to have mastered the knowledge and skills which are the intended performance objectives for the course, should perform at very high levels of mastery on each test task or item. Thus, one determines item difficulties by administration of each item to such groups and recording the frequency of correct and incorrect responses to each group. Items which discriminate between naive persons and skilled persons should be clearly apparent. The naive group should have very high frequencies of error to each item and the expert group very high levels of correct or mastery level performance. If these results are not observed, the test items or tasks need to be modified.

It should be noted that this is precisely the procedure suggested in earlier chapters as a method of external validation of a performance task. The same procedure is carried out, except that data are tabulated on each item for both groups. Items which do not discriminate between the naive and expert groups are eliminated or modified to insure that they do discriminate. Under criterion referenced testing procedures and mastery learning approaches, pre-tests on course knowledge and skill ought to have very high indices of item difficulty when calculated in the traditional manner. Post tests or delayed post tests ought to have very low item difficulties when calculated in the traditional manner.

A discrimination index is typically computed for each item under the norm referenced approach. This, as with the difficulty index, is based on a comparison of scores of persons in the top and bottom extremes of the observed range of scores. A rule of thumb for calculation of a simple item discrimination index for each item is to subtract from the proportion of high scorers in the top third of the total test score and who got a given item correct, the proportion of low scorers in the bottom third of the total test score and who also got that item correct. This is pepeated for every item. That is:

NR Item Discrimination Index = Proportion of high scorers who were correct — Proportion of low scorers who were correct

If an item discriminates well it has a high and positive value approaching the limit of +1. If an item discriminates not at all of poorly it approaches an index of 0. If an item is so poor it discriminates in a reverse manner such that persons with high scores on the total test consistently get the item wrong while persons with low scores on the total test consistently get it right, the discrimination index approaches annegative value of -1.

Again, in criterion referenced testing approaches, discrimination indices are more frequently determined from examination of the results of naive groups of persons' performances on given items, and the performances on the same items by groups of persons expert in the course content. If an item discriminates well the naive group should consistently get the item wrong and the expert group should consistently get the item correct.

Calculation of Item Difficulty and Discrimination Indices for Criterion Referenced Tests

Actual indices of difficulty and discrimination of items may be calculated for criterion referenced tests by means similar to the traditional methods. In the case of item difficulty one can add the proportion of experts who got an item correct to the proportion of naive persons who got the same item correct and divide by two. That is:

$$\text{CR Item Difficulty Index} = \frac{\text{Proportion of experts with correct answer to an item} + \text{Proportion of naive persons with correct answer to an item}}{2}$$

Once again it can be seen that the optimum average item difficulty for the entire test is .5. This value indicates that, on the average, the naive group of uninstructed persons was not able to perform the test tasks but that persons known to be expert in the course knowledge and skill were able to perform consistently at a mastery level. If one develops such a test it can be used to measure the effectiveness of instruction of a given course in bringing the students enrolled in the course to levels of mastery. In addition, within the limits of error of measurement and the validity of the test tasks, each person who has completed the course may be certifed as having mastered or not mastered the knowledge and skills which were the intended performance outcomes for the course.

In a similar manner the traditional way of calculation of item discrimination indices may be modified to calculate a discrimination index for each item for a criterion referenced approach. In this case

CR Item Discrimination Index = Proportion of experts who had an item correct − Proportions of naive persons who have an item correct

If an item discriminates perfectly between these two groups it will have a value of +1. All the experts will correctly perform on the item and all of the naive persons will perform incorrectly. If the proportion of persons in each group is the same in terms of correct performance, the item will not discriminate at all and the value will be zero. This can happen if the item is too difficult so that everyone in both groups gets it wrong or if the item is so easy everyone gets it correct. It can also happen if the item is unrelated or invalid with respect to the content and skill of the performance domain. In such a case, correct and incorrect answers might be randomly distributed across the expert and naive groups in equal proportions. Again, the item would not discriminate between the groups. Therefore, the item should be removed or modified so that it will discriminate.

As in the case of the norm referenced approach to testing, it is possible to have items which discriminate in a reverse manner where the expert group performs consistently incorrectly and the naive group performs consistently correctly. Such an item is poor and will have a discrimination index approaching the value of -1.

Item Analysis Procedures in Perspective

In actual practice as tests and test items are developed it is a good idea to calculate both the traditional norm referenced and the criterion referenced indices of difficulty and discrimination. Detailed procedures more sophisticated than those presented here may be found in Maratuza (1977), particularly in Chapter 17 which deals with criterion referenced testing. There are also many existing computer programs for the routine calculation of these values along with estimates of test reliability. The values which are obtained from such analyses of test items are useful in revising tests to be more valid and to better discriminate between those persons who have learned the intended knowledge and skill taught in a course and those who have not. Should the director of a continuing engineering education program desire assistance in these matters, expert help is usually available in most university testing, counseling, and computing centers. Persons working in these centers are usually facile in these procedures. Other persons with high levels of expertise in this area of methodology are typically found in educational psychology, psychology, and behavioral science departments.

It should be pointed out that the actual calculation of item difficulties and discrimination indices along with test reliability estimates is useful in the task of making better test tasks and items. However, these procedures are not useful by themselves. The procedures outlined in Chapter 10 about how to define performance objectives, sample instructional tasks within these performance domains, and conduct external validation of these instructional test tasks are more basic and prerequisite to good tests than are formal item analysis and test reliability studies. Furthermore, all of these earlier procedures are best carried out by the persons who design and teach the courses with little assistance from persons expert in test construction. Serious attention to the design of good test items and tasks in these other areas does much to insure that the tests which are developed will be sound.

Methods of Reliability Estimation; The NR and CR Cases

The same problem exists for the calculation of test reliabilities as exists for the calculation of item difficulty and discrimination indices. Most of the traditional procedures are based on norm referenced testing procedures and were developed in the interest of producing highly reliable standardized achievement tests. As in the case of item difficulty and discrimination indices estimates, the traditional procedures need to be modified when calculating the reliability estimates of criterion referenced tests. In this section the traditional procedures will first be described and the modifications necessary for criterion referenced testing approaches will follow.

There are four common methods of estimating the reliability of a traditional norm referenced test. These are the alternate forms method, the test-retest method, the subdivided test method, and internal consistency methods (Nunnally, 1972). All of these methods provide estimates of the stability or consistency of the mental measurement achieved by the total test score for individuals.

Alternate Forms Method

The alternate forms method requires the existence of two or more forms of the same test. All forms should be parallel to one another with respect to the knowledge and skill being measured. The reliability of the test(s) is estimated based on the correlation of the scores of the same individuals on alternate forms of the test. The procedure requires the same group of persons to take both forms of the test, preferably at the same time to avoid changes in scores due to experience or other factors. Each person's score is determined on each of the two or more tests. The correlation coefficient is computed based on pairs of persons' scores across alternate forms of the test.

The alternate forms method of estimation of test reliability is a very good method for norm referenced tests. It measures the sources of reliability related to the errors in the sampling of test items for the two tests from the large knowledge domain of interest. It also measures the reliability of the tests in relation to errors due to the fluctuations of individuals' performances over subsequent administrations of the test.

For a criterion referenced test, the alternate forms method is not a good procedure. This is particularly so if the criterion referenced test is designed to determine mastery level of the content area following instruction, which is usually the case. The problem is that both of the alternate forms of the criterion referenced mastery test will show very little variation across persons who have completed the course and mastered the content. Correlation procedures are based upon fitting a line to a set of paired coordinates in order to obtain information about the relationship of one set of scores to the other set of scores. Variation in test scores is required for this to be a meaningful procedure.

Suppose three alternate parallel forms of a criterion referenced test were developed for a short course for engineers. Test forms A and B were administered to six engineers after they had completed the short course. Form A was given immediately after the course and form B was administered the next day. Form C was given as a pre-test before the course was underway. The results of these administrations are shown in Table 6. The means and standard deviations for each of the three tests are also presented. In addition, the observed minimum, maximum, and average mastery levels of the participants are presented for each test administration.

The maximum score which can be acquired on any test form is 60 raw score points. Inspection of the scores of individuals on forms A and B administered as post tests reveals that persons are performing at high levels of mastery. In fact, the lowest observed post test score of 53 is at the 88.3 percent mastery level on form B. On form A the lowest observed score is 55 at a 91.7 percent mastery level. The average mastery level on form A is 95.8 percent and for form B, 94.2 percent. Clearly, from a logical standpoint, two tests have been developed which function very nearly the same. Both indicate, to a high common degree, that the six persons completing the course have mastered the material. Further support for this conclusion is found in the pre-test results for form C. The pre-test scores are those of a naive group, not yet instructed in the course content. These scores are very low, at or below a 35 percent mastery level in all cases. Suppose both form A and form B were used

Table 6

Scores of Engineers on Alternate Forms of Three Criterion Referenced Mastery Tests*

Person	Form A Post Test Score	Form B Post Test Score	Form C Pre-Test Score
1	56	60	15
2	60	53	12
3	59	58	18
4	55	53	10
5	57	60	21
6	58	55	7
Mean Raw Score	57.50	56.50	13.83
Standard Deviation	1.87	3.27	5.19
Minimum Mastery (%)	91.7%	88.3%	11.7%
Average Mastery (%)	95.8%	94.2%	23.1%
Maximum Mastery (%)	100.0%	100.0%	35.0%

*All scores reported in raw score units unless otherwise noted. Maximum possible score on any form of the test is 60.

For the relationship $y = mx + b$ for Form A and B scores, $m = -0.31$, $b = 74.57$, and the correlation coefficient between forms A and B is -0.18.

as a pre-test for other groups and that the results obtained were very much the same as those obtained when form C is used as a pre-test. Suppose that form C was also used as a post test for some of the other groups and, once again, the results obtained were very similar to the post test results for forms A and B and for the present example. On the basis of this information it would be reasonable to conclude that the forms A, B, and C as post tests are reliable. They produce highly similar and replicable results. This conclusion is strengthened if the three forms of the test continue to be used in the post test role with other groups of persons completing the course and these groups also show consistently high levels of mastery of the test material.

Yet if the reliability of form A is calculated based on correlation of the scores of the six persons with form B using the data in Table 6, a very different conclusion would be reached. The correlation coefficient for the two sets of scores may be calculated from the slope of the line fitted to the six sets of paired data points by the equation,

$$y = mx + b$$

The correlation coefficient can be found by the relationship

$$r = m \frac{s_x}{s_y}$$

where:

r = the correlation coefficient

m = the slope of the fitted line

s_x = the standard deviation of the values in the x array of scores

s_y = the standard deviation of the values in the y array of scores

Using this relationship and the information contained in Table 6, the reliability, in the form of the correlation coefficient between two alternate forms of the test, may be estimated for the tests developed for the short course. When the proper values are substituted into the above equation, after the slope has been determined and the standard deviations of both sets of test scores have been calculated, the following results are obtained.

$$\hat{r} = -.314 \frac{1.871}{3.271}$$

$$\hat{r} = 0.180$$

The reliability estimate for the test forms is very low. Under any usual norm referenced testing procedure the test would be judged as being very unreliable and would be discarded.

The basic problem is that this method of estimation of reliability was developed for situations where there is a very large domain of material for which a huge number of items may be written. In such cases it is usually of interest to sample some items from within the domain by which to estimate an individual's total knowledge of the broad domain. It is also of interest to rank individuals in some population of persons in terms of how much knowledge of the domain each individual in the sample has with respect to other persons in the population. Under such circumstances persons can be expected to vary quite a bit in their scores on the test. The correlation of alternate forms of a test in making estimates of individual persons' knowledge domain is the matter of interest. The variation in test scores between persons in the sample is required for alternate test form reliability estimation procedures. Typical achievement tests of a comprehensive nature are good examples in which such reliability estimation procedures are appropriate. This is because the knowledge domain being tested is very large and the inclusive sample of persons used to norm the test varies greatly in ability levels. Therefore there is much between persons variance in the test scores of individuals, even though with a highly reliable test any given person's test score will not vary much over repeated administrations of the test.

However, in the situations presented in this book it is clear that the domain of knowledge or skill to be tested is usually much more circumscribed. Usually the question of interest is, "How well have the course participants learned to do the specific things the course was intended to teach them to do proficiently?" The interest is in producing instruction in areas of complex knowledge and skill which lead to uniform high levels of consistent performance across persons, and in achieving mastery of a relatively small number of specific concepts, skills, or procedures which are the goals of the course.

In addressing this problem, it has frequently been suggested the best form of reliability estimation for criterion referenced tests of a mastery level type is the simple degree to which the results replicate from one test to another and especially from one group of persons who have completed the training to other similar groups who have also completed the training (Maratuza, 1977, Chapter 12, Tyler, 1974). Of course, one must be sure that the tests developed do consistently discriminate among naive and expert groups of persons before making the inference that replication of results of a given post test with multiple groups of trainees after course completion means that learning has occurred. One can also accomplish this procedure by randomly assigning a group of enrollees to a pre-test administered before instruction and the remaining half to a post test administered after instruction. Comparison of the scores of these two groups reveals if the test is functioning appropriately without confounding the interpretation with repeated measurement of the same individuals with similar test forms (Maratuza, 1977). Without the base line data on the test for naive groups, one does not know whether the performance on the post test reflects learning resulting from the course or learning acquired before the course. This is another reason for the use of parallel comprehensive pre- and post tests in the development and evaluation of short courses, as was suggested in earlier chapters. These tests are also useful for making estimates of learning outcomes for individuals resulting from such courses, and these measurements may be reported as estimates of individuals' learning.

Test Re-Test Method

Another common method of estimating the reliability of a test is the test-retest method. This involves administration of the same test form twice to the same group of persons with some period of time between administrations of the two forms to prevent memory from playing a key role in the production of the second set of responses. Again, the scores from the two administrations are correlated across persons and the correlation coefficient obtained is used as an estimate of the reliability of the test. The same problems described earlier in relation to the alternate forms procedure as a method of reliability estimation for criterion referenced tests apply here as well. If persons have mastered a set of particular skills, concepts, and procedures, there will be very little variation in scores from one test administration to another and from one person to another. The correlation coefficient will not beaa good estimate of the test reliability. There are also practical problems involved in administering the same test twice for a short course where the total number of items and the scope of the content covered in the course may make improved performance on the second administration, because of practice and memory factors, more likely than on a test consisting of many items sampled from a very large and general domain of knowledge and skill.

A modification which is appropriate to reliability estimation for criterion referenced tests does exist (Millman, 1974). Under this procedure one develops a large pool of items for the area to be tested, with attention to having parallel forms of items. Next, two test forms are produced by random assignment of the items or item pairs to test forms. These items are then intermingled into one test and the test is administered to the persons who have completed the short course or to some other groups such as an expert or naive group. The test is then scored. The score for each person on form A and form B is determined. A chart or graph is then prepared which lists the absolute difference in scores of each person on both forms of the test. Inspection of the results reveals how parallel the two forms of the tests are and how reliably they measure the domain of specific knowledge and skill which is of interest. An additional procedure is to calculate the mean value of the absolute difference score across persons. This value serves as an index of the degree of consistency to which the two test forms measure persons' competencies in the area tested. The closer the mean difference value is to zero, the more consistent are the tests. This method is not affected by a lack of variability among persons on the test scores because of achievement of mastery. It is, however, affected by the difficulty levels of the items. Therefore, information about the test difficulty level must be considered in making a judgment of the reliability by this method. It is also clear that the most appropriate use of this method is with a group of experts who are known to have mastered the knowledge, skills, and procedures of a particular short course, or persons who have been taught to master this body of material by having completed the course successfully. The reliability of the test is of interest in relation to the test's ability to consistently discriminate between persons who have mastered and are expert in the content of the course versus those persons who have not and are naive.

The similarity of this procedure to the suggestions made in earlier chapters for the development and use of pre-tests, embedded tests, post tests, and delayed post tests should be obvious. In all cases the matter of primary interest is to determine if course objectives have been achieved by students and to insure that tests are developed which are capable of answering this question. It is of little interest to rank order persons in terms of the degree to which they comprehend some complex and inclusive field of general knowledge. It is of great interest to determine if a given course is resulting in students learning specific skills and procedures to mastery levels.

Subdivided Test Method

Another traditional method of estimating reliability of a test is to subdivide the test into two parts. This is often done by considering all of the odd numbered items to consist of one form and the even numbered items another form. The test is then administered to a group of persons. Each person's test is scored twice, once on the odd items and once on the even items. The pairs of scores for individuals are correlated and the correlation coefficient is taken as an estimate of the reliability of the test. The procedure is similar to the alternate form and test-retest methods. The same limitations for criterion referenced tests exist for this procedure as in the other two cases. The same methods for overcoming these problems, as described by Millman (1974) and others, also apply.

Internal Consistency Methods

Perhaps the most common method of estimation of the reliability of a norm referenced test is the Kuder-Richardson formula 20 method, commonly referred to as the KR 20. The basic rationale for this method is that if the test is internally consistent, all items should tend to measure the same thing and should correlate highly with one another. If all the items in a test correlate highly with one another, the inference is that the test would likely correlate highly with an alternate form which does not exist but which would be composed of similar items.

There are two common forms of the KR 20 formula. The first form is used to calculate the reliability of a test where each item is scored as being wrong or right. The scoring for each item must be dichotomous, as is usually the case with multiple choice tests where every item is scored as correct or incorrect. This formula is:

$$r = \frac{n}{n-1}\left(1 - \frac{\sum_{i=1}^{n} pq}{s_t^2}\right)$$

where: r = reliability estimate for the test

n = number of items on the test

s_t^2 = observed squared standard deviation or variance of the total test scores across persons

p = proportion of persons passing a given item

q = proportion of persons not passing a given item

Another version of the formula is available for test items where partial credit may be awarded rather than a 0 or 1 score only. For example, if one were to administer a test consisting of 10 problems where each completed problem is scaled from 0 to 10 points in terms of completeness and accuracy, one would use the second form of the KR 20. Here the formula is:

$$r = \frac{n}{n - 1}\left(1 - \frac{\sum_{i = 1}^{n} s_i^2}{s_t^2}\right)$$

where: r = reliability estimate of the test

n = number of items on the test

s_t^2 = observed squared standard deviation or variance of the total test scores across persons

s_i^2 = observed squared standard deviation or variance of each individual item score across persons

It is obvious from inspection of the two formulas that the problems noted earlier about the use of methods of estimation of reliability of norm referenced tests for criterion referenced tests also apply here. The reason is that once again, the procedure depends upon having a large variation in performance across individuals on test items. The almost certain expectation is that the variation will be too low in cases of criterion referenced tests, under conditions of mastery learning. Consequently, the KR 20 method is inappropriate for estimating the reliability of a criterion referenced test under mastery learning conditions.

As in the other situations, there exist methods by which to modify the KR 20 procedure for use with criterion referenced tests of mastery level (Hambleton & Novick, 1973; Livingston, 1972, 1973). One way is to deviate each person's score from an arbitaarily determined mastery level rather than from the group mean score. This deviation can be used instead of the usual variance values in the KR 20. However, this method still requires a sufficient amount of variation across persons on test items and in their total test scores. The reliability estimates of test results similar to those listed in Table 6 by this method would still yield low values. The best method of reliability determination for criterion referenced tests is consistent replication of results across repeated testings of similar naive and expert groups, as has been described earlier.

Conclusion

Attempts should be made to determine the reliability of tests used in continuing engineering education courses even when these tests are of the criterion referenced type and mastery learning is the expectation. This will usually be the case in most short courses. An exception is the case of the short course concerned with remediation of persons' knowledge and skill in a very broad domain, as is the case in courses which prepare engineers to take licensing or certification examinations. Here there is a large and broad domain of knowledge and skill involved. There are norms determined by the test scores of all certified professional engineers on these examinations. In this case, it is of interest to rank order persons and to determine something of an individual's knowledge of the broad domain with respect to other engineers in his or her specialty area. In this situation, the tests which are developed for assessment of the learning outcomes of remediation or preparatory courses are most appropriately constructed from items sampled from across the very large domain of knowledge and skill under consideration. The pre-tests, which may be used to advise students whether or not they need to take a remediation course, and the post tests, which measure the degree of learning resulting from completion of the course, are most appropriately shorter but parallel to the achievement tests used by the professional engineering societies for licensure purposes. In this case, all of the traditional norm referenced procedures for the calculation of item difficulties, discrimination indices, and test reliability estimates are very useful and highly appropriate.

There may be other times when the goals of a particular short course are more concerned with changing and improving some broad area of knowledge rather than producing highly specific skill and knowledge outcomes. In these cases, the courses may be more like the typical remediation or preparatory course and it may be appropriate to use norm referenced testing procedures by which to develop, validate, and determine the reliability of the tests. However, most short courses in engineering for continuing education purposes will by necessity be directed toward a much more defined set of specific intended performance capabilities which should be achieved to high levels of mastery following instruction.

Therefore, the criterion referenced test approach to matters of determination of item difficulties and discrimination indices and to estimation of test reliability will generally be necessary.

The procedures outlined here for the modification of the usual norm referenced testing approach to meet the needs of criterion referenced and mastery learning approaches are very elementary. For more information on how to accomplish well designed reliability studies the reader is advised to refer to Applying Norm-Referenced and Criterion-Referenced Measurement in Education by Maratuza (1977). This book is particularly useful since it presents both the traditional procedures and detailed explanations of how these procedures need to be modified for the case of criterion referenced testing and mastery learning expectations. Another source, which is widely used as a guide to the carrying out of item analysis and test validity and reliability studies, is Educational Measurement and Evaluation by Nunnally (1972). This latter book makes no reference to the criterion referenced testing situation but provides all the basic information and procedures for the traditional norm referenced situation.

Once item analysis and test reliability studies have been completed, it is usually necessary to modify some test items and to revise whole tests in order to produce better measuring instruments. The modifications and adjustments necessary should be made with reference to the matrices and plans described in Chapter 10, which produced the original items. The data from formal item analysis and reliability studies is not by itself sufficient to the task of rewriting and revising individual items and tests. It is only useful in pointing out difficulties and problems with particular items and test forms which might not otherwise be noticed. The actual business of rewriting a given test item to better discriminate between naive and expert persons, or the actual modification of a test for the same purposes, is basically a logical task dependent upon the content of the course of instruction, the intended performance outcomes, and the matrix of these as they define the purpose and intent of instruction. The procedures outlined in Chapter 10 are the basis for not only the initial development of test items and tests, but also for the revision of items and tests to better serve their intended discriminative functions. Without attention to the procedures outlined in Chapter 10, the item analysis and test reliability procedures described in this chapter have little utility or meaning.

Even well designed tests have inherent limitations as estimates of learning outcomes. These limitations and their implications for the proper use of tests are the topics of the next chapter.

Chapter 12

LIMITATIONS OF TESTS

The limitations noted in this chapter apply to traditional norm referenced testing procedures in particular. Criterion referenced testing procedures, similar to those advocated in earlier chapters, are less subject to these limitations. However, there are limitations for any procedure which uses test results to make inferences about the degree of an individual student's learning resulting from a course of instruction.

Sources of Invalidity and Unreliability

No test, no matter how well constructed, is perfectly valid or reliable. This means there is always some question about what the test measures in relation to the domain of performance and knowledge it is supposed to measure. There is also another question about the consistency with which the test measures whatever it is that is being measured. Because of these questions all test scores have elements of error in their estimates of the performance capabilities of persons. Even if a population of persons could be found in which every person had a precise and unchanging amount of knowledge and skill in the performance domain being tested, repeated administration of even a very good test to the same group would produce variation in test scores of these persons.

This variation is because: 1) there are always multiple ways in which a given test item may be interpreted by the persons completing the task; 2) the way persons respond to a given test item or task are subject to influences such as time of day, degree of hunger, the presence or absence of personal concerns and worries, and many more uncontrolled variables which cause performance to vary; and 3) there are often variations in the way test item responses are scored or judged "correct" or "incorrect".

With objectively scored tests, variation in scoring or judging the degree of "correctness" of persons' responses is removed. Objectively scored tests consist of tests which have a scoring key by which to determine unambiguously if a given response to a given item is right or wrong. One example of objectively scored tests is the multiple choice test format which is widely used because of this property. However, even an objectively scored test is not truly objective. There are still the usual ambiguities and idiosyncratic variations in each item which cause the test to be less than perfectly valid or reliable, no matter how objective the scoring procedure.

Even when test items are problems of an engineering nature requiring the use of particular course concepts and skills to set up and work a problem to determine the specifications of a particular piece of equipment or obtain certain numerical values, there is a great variation in the scoring of the results. In a study involving 1,071 professors of engineering, mathematics, and physics, Clyde Work (1976) found huge variations in the scoring of a set of common student responses to eight questions on an engineering statics and dynamics examination. The responses were actual responses of students to a real examination. The persons who scored the examination were a national sample of professors, from universities and technical colleges, who actually taught such courses. Each item was corrected on a 0 to 10 point scale where points were deducted for errors up to a maximum of 10 for any given item. The results showed that for most items, the scoring variations ranged from no points deducted to all 10 points deducted! For most other items the spread in points assigned to individual items was also very large, in most cases exceeding six points. It should be recalled that what was being scored by these professors were the same answers to the same problems by the same students. The variation in scoring was caused by the interpretations of individual professors. Some took off points for lack of neatness. Some did not. Others gave variable amounts of partial credit for correctly setting up a problem while others did not. The net effect of all of these individualsjudgments by professors about the accuracy of the student's response is a huge variation in the performance score assigned the student. On any given problem the same student response is likely to be graded by different professors all the way from completely wrong to completely correct with all intermediate scores being well represented (Work, 1976).

Other reasons for validity and reliability problems of tests have to do with the abbreviated and incomplete sample of performance which is encapsulated in each test item. Test tasks must always be abbreviated, shortened, or otherwise reduced in complexity and time demands in order to allow their completion in a reasonable length of time. This means that test tasks are artificial samples of the domain of performance which is of interest. The manner in which the tasks are sampled from the actual domain of performance and the manner in which they are abbreviated directly effect the validity of the test tasks. Even when test tasks are abbreviated in the best manner possible, there remain problems in inferring from the test tasks the ability of the person in real world performance tasks which are typically more complex, carried out over a longer period of time, and typically accomplished with the help and support of other persons in a cooperative manner. For example cooperation on a test task is considered cheating and is usually punished. Yet cooperation in carrying out the solution to complex problems in real work activities of engineers is highly valued and encouraged.

For all of these reasons, making inferences about the actual capabilities of persons in areas of complex skill and performance on the basis of a test score is risky. Test scores are more useful for providing information about the degree to which basic information, concepts, relationships, and procedures have been learned and understood than for making predictions about who will actually perform well on the job in dealing with a complex set of tasks. Even well designed standardized tests, such as are frequently used in admissions procedures in colleges, as well as the standardized achievement tests used to certify persons competent in an academic discipline, do not predict adult achievement in actual work settings (McClelland, 1973; Stice, 1979). These types of test scores do predict academic success of students in formal college and university programs, but only to a small degree. Standardized achievement tests, even when matched directly to the content of academic programs, at best account for only 25 percent of the variance observed in student achievement (Lavin, 1965). The other 75 percent of the variance in observed student achievement in academic courses and programs is attributable to differences among students in interests, motivation, persistence, experience, and opportunity.

Scores from tests developed for courses and group statistics are more accurate predictors of the success of a given course in teaching certain concepts, information, and procedures to persons on a regular basis than they are as precise measures of an individual's learning. This is because the effectiveness of a course can be determined by using the test data across many individuals who have participated in the course and completed its tests. The average data is much more stable than any individual test score because idiosyncratic responses of persons to test items tend to balance one another out. The standard error of the mean score of a group of persons who have been tested is always much smaller than the standard error of the estimate of an individual's test score.

The Standard Error of Estimate

The standard error of an individual's test score is an estimate of the deviation of a person's particular score from that person's true score on a given administration of a test among many repeated administrations. Of course, it is not possible to repeatedly test one person for as many as 15 to 20 times on a given test. It is also not possible to assume that a person would not change in his or her knowledge over repeated administrations of the same test. However, in an imaginary situation, if a person were tested repeatedly on one test and the standard deviation of his or her score were calculated, a standard error of the estimate for the score of that person would have been calculated. The person's true score would be estimated by the mean score of the repeated administrations.

In practice the standard error of estimate of a test score is calculated in other simple ways. One common procedure for the calculation is listed below.

$$\sigma_{measure} = \sigma_t \sqrt{1 - r}$$

Where σ measure = standard error of measurement of the test

σ_t = standard deviation of the test for the sample of persons completing it

r_t = the reliability of the test obtained for the sample of persons completing it

The standard error of measurement for a test provides information about a confidence interval or band around each person's observed test score. The greater the reliability of the test, the smaller the confidence interval and the more accurately a person's observed score predicts the person's true score. The standard error of a test score allows the assignment of odds to the liklihood that any person's observed score is actually different from some other person's score versus being different from each other only due to chance factors resulting from the less than perfect reliability of the test. Perhaps an example will help.

Suppose a 20 item test is developed for a continuing engineering education course in soil mechanics. The reliability of the test is calculated by the KR 20 alpha coefficient method to be .74 based upon a sample of 26 persons who took the test. The raw score test mean and standard deviation for this sample are 13.08 and 3.77, respectively.

The standard error of measurement for this test would be calculated by substitution of the correct values into the equation listed above. The results would be:

$$\sigma_{\text{measurement}} = 3.77\sqrt{1 - 0.74}$$

$$\sigma_m = 1.92 \text{ raw score units}$$

If we wished to be correct 95 percent of the time in classifying persons as having scored differently from one another or from some arbitrary level of performance, we would have to multiply 1.92 times 1.96 to obtain a confidence band around an individual's observed score. Suppose Mr. Perkins, a student in the course, obtained a score of 12 on the test. Multiplying the standard error of the test times the value of 1.96, the number of standard deviations both sides of the mean on the unit normal curve inclusive of the 95 percent confidence interval, a value of 3.76 is obtained. This means Mr. Perkins' true score on the test will fall within 12 ± 3.76 raw score units on this test 95 percent of the time over repeated administrations of the test to Mr. Perkins at this point in his ability with respect to course knowledge and skills. Suppose the course instructor had set a score of 15 raw score points as the maximum mastery level for which successful completion of the course would be recognized. Could Mr. Perkin's have performed adequately and the difference between his observed score and the criterion score be due to error of measurement and unreliability of the test? The answer is, "yes". At his present ability level, Mr. Perkins' true score on the test could be expected to fall within the observed values of from 8.24 to 15.76 on repeated administrations of the test 95 percent of the time.

Stability of Group Mean Scores

The values presented for the reliability of the 20 item test and the standard error of measurement for the test are rather typical for such tests. The example makes it clear that even with good tests, error of measurement of an individual's ability is large. However, the error contained in the test score estimate of the average ability of persons completing the course is much smaller than the standard error of estimate of an individual's test score. Group means are very stable and the standard error of the mean, or average score of a given group of students, is very small. This makes it possible to accurately estimate the effectiveness of a course in achieving its intended learning outcomes based on the average performance of participants on post or delayed post tests. If the tests have been carefully developed following the procedures suggested in earlier chapters, this estimate of course effectiveness will be quite valid as well as accurate and consistent. The estimate is simply how typically effective the course is in achieving the specific knowledge and skill outcomes it posits as its specific performance objectives. The estimate is not about the ultimate value or effectiveness of the course in real world performance of the engineer in his or her work setting. It is a well known fact that course grades and test scores do not predict adult achievement in real world work settings (Hoyt, 1965; McClelland, 1973; Stice, 1979). What they do report is how well specific course content and skills have been learned and, to a limited extent, how well the student is able to learn more related course content and skills in future courses. Using both past course grades and well designed standardized achievement tests as predictors of future academic performance, the maximum amount of variance which can be accounted for in actual student performance is approximately 25 percent (Lavin, 1965). One should always keep these limitations in mind when the results of continuing engineering education students' performances are being reported to them and their employers.

In summary, estimating an individual's knowledge and skill acquisition following completion of a course and based on a single test score is a much less accurate procedure than estimating the average effectiveness of a course in achieving its knowledge and skill objectives based on many students' scores on one test. Test scores pooled across persons can be valid indicators of course effectiveness. Test scores alone for individual students are less valid indicators of the achieved level of learning outcomes. In short, tests are more useful for evaluating courses than for evaluating persons.

In making decisions and inferences about the amount of learning which has been achieved by individuals, multiple and independent test scores are more useful than a single test score. However, the inference about degree of learning should not be based only upon test scores, even if multiple tests are used. What is needed in addition are observations of students' performances on complex tasks and careful examination of the products produced by students in the execution of complex performance tasks which have been assigned. As mentioned in earlier chapters, these products of performance include designs, analyses, reports, solutions to complex problems, plans, and other products normal to the work performance of the engineer and upon which the content and skills of a particular course can be brought to bear.

Other sources of information about the degree of learning of course content and skills should not be overlooked. These include asking the course participant how much he or she has learned, how useful the learning was in relation to improving work performance in particular areas, and how often course content and skills are actually being used on the job. Supervisors and employers should also be asked to make similar judgments. If all of this information is collected, it is much more appropriate to make an inference about the degree of learning of an individual than it is to do so based only on a single test score or even multiple test scores.

Advantages of Criterion Referenced Tests

Although many of the limitations of tests also apply to criterion referenced tests, the restrictions are not as great. There are a number of reasons for this.

Traditional norm referenced tests tend to be global and not well defined in terms of the particular knowledge, skill, and performance capabilities they are testing. Consequently, the items selected for inclusion on the test are only a few sampled from a huge domain of potential items. This is the normal situation in standardized achievement tests. In criterion referenced tests the focus of the testing is on very specific and well defined knowledges, skills, or performances not in comparing the performance of individual persons on the test against other persons in a sample of similar persons. The student's performance of the specific tasks presented in the test is of interest because these tasks are the intended goals of some preceding instructional activity which is specifically designed to teach that performance.

The criterion referenced testing situation is more tightly controlled than is the typical norm referenced testing situation. On a criterion referenced test the student demonstrates what he or she can do on some specific task which has been taught. Typically the student will have to recall knowledge and information learned, make judgments about how to proceed, and apply skills learned to complete the task. After the task is completed by the student the course instructor can determine how well it was completed. The product of the performance can be compared with the product of the performance of a skilled professional for the same task. The scoring of the accuracy and the completeness of the student response can be quite objective with respect to some standard criterion of acceptable performance.

In a typical norm referenced testing situation there is no comparable grounding of the test items or tasks in specific aspects of skilled performance to a criterion level. Thus individual test scores are not directly interpretable in terms of what the student can or cannot do in the way of executing a complex performance with skill. All that can be said is how well the student has performed on a set of test tasks in comparison to a group of peers. The judgment of the student's performance capability is relative and general. In the criterion referenced case, the judgment of the student's performance is absolute and specific to well defined skills or capabilities.

If properly developed and administered, criterion referenced tests have an inherent validity and reliability which makes them superior to norm referenced tests (Tyler, 1974) particularly in the content of short courses and other learning experiences typical of many continuing education courses where what is to be instructed and what is to be achieved by students are very well defined and highly specific.

There can still be problems of agreement among different raters regarding a student's performance on a criterion referenced test task. Just as Work (1976) noted wide variation among instructors who scored the same student's answers to a common problem, instructors scoring the performance of a student on a criterion test item can disagree. However, they would not typically be expected to disagree as much because part of developing a criterion referenced test is the determination of what will be taken as evidence of correct performance. In short, a common, high level of criterion performance is defined before testing, and even before instruction. Consequently, during a course students are specifically instructed and guided toward exhibiting mastery of the knowledge and skills being studied. It is clear both to them and to the instructor what it is which is to be accomplished and how well the instructional and test tasks must be performed. There is a common and clearly communicated agreement on the standards of performance which are acceptable. These criteria are usually grounded in the standards that are judged acceptable in actual practice of expert engineers on similar tasks in actual work settings.[1]

---

[1] The education of professional persons to high levels of mastery in complex skills and abilities is typically approached in this manner. For a detailed logical, philosophical, and empirical exploration of this topic the reader is referred to a study by Lacefield (1980) concerning the measurement of competence.

Thus it is reasonable to infer whether or not a student, having completed a course of instruction, and having performed on a criterion referenced test, has achieved the desired level of competence or skill. Criterion referenced performance assessment is used particularly in situations where it is critical that the student actually be competent before being allowed to practice. As was mentioned in an earlier chapter, these types of test tasks are used with physicians before they are allowed to conduct surgery, with aircraft pilots before they are allowed to fly planes, and with persons who operate very expensive and complex equipment where an error or lack of competence on the individual's part would be a threat to property and life. In such cases criterion referenced performance tasks are routinely used to conduct assessments of the individual's competence. The test tasks are very specific to the performance capability under consideration. They are very valid and reliable because they involve simulations or abbreviated test tasks which demand nearly the same types and quality of performance as does the actual performance.

The procedures outlined in Chapter 10, which describe how to develop good assessment procedures and test tasks by which to measure learning outcomes for continuing education courses, are directed toward producing criterion referenced tests. The procedures in Chapter 11 explain how traditional norm referenced testing procedures for determining item difficulties, discrimination indices, and test reliability need to be modified for criterion referenced tests. Use of the traditional item and test analysis procedures for criterion referenced tests is inappropriate. Just as the traditional concepts of test reliability do not apply to criterion referenced tests, neither does the traditional concept of standard error of measurement of a test score.

In Chapter 11 it was noted that all of the usual methods of estimation of the reliability of a test depend upon the occurrence of a large amount of variation among the test scores of students completing the test. If there is little variance among scores the reliability will be low and the procedure for computation of the test reliability will be invalid. The same problem exists even with methods such as Livingston's (1972) which is designed to modify traditional procedures of internal consistency reliability estimation for criterion referenced testing. There often is simply too little variation in student scores upon completion of the course for these procedures to work well.

When course objectives have been clearly laid out, when instructors work to insure that the student has mastered the course content, and when nearly all students upon completing the course have mastered the material as judged by their test performance, then it is clear much learning has occurred. Many studies show that instruction of this type is superior to traditional instruction in terms of the amount students learn, how long they retain it, and how well they can apply it in the next course (Kulik & Kulik, 1976; Kulik, et al., 1979). Yet if one were to compute the reliability of the tests used to make determinations of the students' achievement in these courses, using the traditional means, the tests would have poor reliability. This is because the variance in test scores is reduced because of uniform high achievement by students. Yet we know that the tests are reliable estimates of students' learning because students who have completed such courses and scored highly on the examinations for them consistently out perform other students taught in traditional courses. The mastery learning students perform better than students instructed in traditional ways on standardized examinations of the content area, on common final comprehensive examinations, and in achievement in future and more advanced courses in the same content area (Kulik & Kulik, 1976). These results are consistently found over many empirical studies conducted in engineering and the physical and social sciences (Kulik,et al., 1979). Clearly the tests must be measuring student achievement of learning outcomes in reliable ways or these consistent results would not hold.

As was pointed out in Chapter 11, the best indication of the reliability of a criterion referenced test is its ability to consistently discriminate between groups of persons known to be skilled in the performance of interest and those who are known not to be skilled. Replicability of results with multiple groups of naive and expert persons is the best indication of test reliability or consistency (Tyler, 1974). The best means to insure this replication of results is to develop the test tasks and the instructional procedures according to the steps outlined in Table 3, Chapter 10.

Likewise, the construct of the standard error of measurement does not apply well to the criterion referenced situation (Livingston, 1973). Other procedures more appropriate to estimating true mastery scores have been developed. These methods are quite different than the concept of a person's true score and the standard error of estimate of a test score which are used in norm referenced testing (Hambleton & Novick, 1973). Rather than compute the probability of a person's observed score falling within some range around his or her true score, in the criterion referenced testing situations one is interested in asserting whether the persons can or cannot perform the task correctly. The probability is 0 to 1. The judgment is accurate only insofar as the test task has been properly grounded in the performance domain of interest as outlined in Chapter 10.

Even criterion referenced tests must necessarily consist of abbreviated performance tasks. Flying a flight simulator on instruments is not really the same as flying a real commercial aircraft in bad weather with a full load of passengers. Designing an automation procedure for a simulated industrial production process and testing the procedure with a demonstration microprocessor kit and a computer program is not the same as actually developing and installing an automated procedure in a real factory. Even the best simulations, abbreviated test tasks, and practice examinations must always be only initial and partial assessments of the individual's probable competence in the actual work domain with its many uncontrolled variables and greater complexity. Yet, the limitations of test tasks, because they are abbreviations of actual work situations, are less for criterion referenced tests than for norm referenced tests. This is because there is an explicit attempt to tie criterion referenced tests directly to the performance required in the actual work domain, and because the capabilities being tested are much more specific and defined than in the norm referenced situation.

Even with criterion referenced testing procedures it is possible to make even stronger inferences about the success of groups of persons achieving the desired learning outcomes following a course of instruction than is the case for individuals. Demonstration of consistently high levels of complex performance by many participants following instruction provides convincing evidence that the course is effective in achieving its intended outcomes.

Summary of Major Points Concerning Testing

The next section of this book deals with reporting measurements of learning outcomes of courses to individuals and groups. Before beginning this topic, a summary of the main points made in this and the previous chapters on testing is presented.

Tests can never measure real performance in true to life situations. They must always consist of abbreviated tasks and samples of activities designed to assess knowledge and skills believed to be basic to effective practice in some area. Properly designed tests can reveal much about how well a course is fostering its intended learning outcomes. Tests have limited value for making inferences about individual persons' performance capabilities in actual job situations. Although these limitations apply to all tests, they are less applicable to properly designed criterion referenced tests.than to typical norm referenced tests. Consequently, tests designed according to the procedures outlined in Chapter 10 and elsewhere (e.g., Maratuza, 1977, Chapter 17) provide much better estimates of individual achievement of intended learning outcomes than do other types of testing. They also provide better estimates of the effectiveness of the course and its instructional procedures.

By stating course objectives in performance terms and by logical sampling of test items and tasks within the domain of performance of interest, it is possible to construct tests which have reasonable validity. External checks on the validity of tests, involving the use of experts to examine and critque the content of the test and to actually complete the test, help to further validate a test. Administration of the test to naive and expert groups will also help develop measures capable of distinguishing between the presence and absence of the knowledge and skill areas of interest.

Item analysis and test reliability studies may be carried out once sufficient data is collected from actual test administrations. These procedures can help refine a test and achieve a proper balance of easy and difficult test items. Adjustments can also be made in the ability of test items to discriminate between persons skilled in the content and knowledge of the course and those who are not skilled or have less skill. Reliability estimates may be obtained in several ways. These estimates are useful for determining to what degree a test is consistent in measuring aspects of a knowledge, skill, and performance domain across individuals and groups.

It is a time consuming task to construct good tests. Even the best tests have less than perfect validity and reliability. There are many reasons for this, including the abbreviated nature of test tasks, the artificial work setting in which they are administered, and the usual variations in human emotion, motivation, attention, and changes in interpretation of the content of test items by different persons and the same persons in different test administrations. However, if tests have been carefully developed, one can make strong inferences about the general effectiveness of a course in teaching specific information, skills, and procedures to groups of participants. This is because the inferences about the group performance and the effectiveness of the course are based upon aggregated data across individuals and result in statistical means for which there is relatively large consistency or stability from one replication of a course to another, all other things being equal. The inference about any particular individual's learning based on a test score is an estimate subject to much more variation.

Consequently, care must be taken in the use of test scores as *the* means of making inferences about the degree of learning achieved by persons as the result of short courses in engineering work performance areas. The best use of test scores for individuals who have completed short courses is to recognize them as rough estimates of the knowledge and skill levels of persons in some specific areas which have been taught because they are believed to be related to the complex performance domain. The most inappropriate use of such scores is to report the individual qualified or unqualified to perform in the actual complex, on-the-job work area especially on the basis of a single test score. Other information based on actual observation of an individual's performance on complex tasks and evaluations of the products of those performances is necessary to make such inferences.

Another good use of test scores is to make inferences about the general effectiveness of a given course for purposes of formative evaluation and subsequent revisions in course content and presentation to better serve the needs of participants. Still another highly appropriate use of test score data is the summative evaluation of the effectiveness of a course at any given time in its history in order to provide prospective clients, course developers, and others with good information about typical course effectiveness in teaching particular skills, knowledge, and information.

Questions about how qualified persons are for job performance in actual work settings after they have completed specific short courses should always be based on multiple indicators of the individual's performance capability in realistic work settings with representative problems or tasks. About all test scores for short courses can reveal is the degree to which the individual engineer has learned knowledge, skills, and procedures thought to be basic to good practice in the actual performance area. High scores do not mean that the person is necessarily competent. Low scores can mean that the person is unlikely to be effective in the performance areas. However, scores lower than arbitrary criterion performance levels can also be caused by measurement error in the test instruments themselves, by inadequate instructional procedures, or by a lack of readiness of the student to enter into and profit from the instructional activities. One should be alert to and attempt to control or compensate for these factors when using test results and other information about performance to make decisions about individuals' capabilities or the effectiveness of courses and programs.

Chapter 13

REPORTING THE ASSESSMENT OF LEARNING OUTCOMES

Persons with interest in the learning outcomes of continuing education courses include the participants enrolled in the course, the faculty who have developed and taught the course, the employers and supervisors of the course enrollees, the administrators and governing bodies of the academic unit involved in operating the course, and the governing bodies of professional engineering societies who are concerned with the quality of the course. For these reasons, there are a number of different purposes and methods for reporting the resulting learning outcomes for any continuing education course. A wide range of information must be collected to meet the needs of these persons to know the effects of a course on enrollees' performance. All of the persons involved need this information for the purpose of decision making (Lacefield, 1980).

Participants' Needs

The individual engineer is most concerned with information about his or her understanding of specific concepts and procedures taught in the course. Information about student success in the specific tasks of the course is part of the instructional process. Both participants and instructors need to know what individual students have and have not yet learned fully and specifically what remains for them to learn more about. The purpose of learning assessments in this context is to facilitate learning by the individual. The results effect subsequent decisions by the individual student concerning what parts of the course to study more, where to ask for assistance, and what future areas of study to engage in on one's own or by participation in additional courses or other formal study activities.

At the end of the course of instruction, the enrollees also have a strong interest in how well they have performed. Information concerning the degree of mastery of course content and skills achieved by individuals is of interest to these persons. People also often want to know how they performed in comparison to others in the group. Instructors should provide both types of information to individual enrollees.

Instructors' Needs

Course instructors have similar information needs in order to carry out the instructional activities and decisions required for the teaching of the course. Decisions about when individual students and groups of students understand course concepts and can go on to the next task require some sort of assessment. This assessment must be made rapidly and during the course of actual instruction. Embedded test tasks such as those described in Chapter 7 are particularly useful to this end. Typical embedded test tasks include quizzes, homework problems, laboratory activities, and other procedures which require the participants to demonstrate their ability to understand and use course concepts and skills in specific ways. This type of assessment activity is closely tied to the business of instruction. It is important that the information collected from such assessments be shared immediately or as soon as possible with the course enrollees by instructors. Instructional decisions concerning pace of activities, provision of individual assistance for students, and prescription of remedial study or work for individuals or groups have their basis in this ongoing assessment of learning.

Course instructors also need to know how the students perform at the end of the course in order to make judgments about the effectiveness of the course for particular individuals and groups compared to other persons and groups. The information about end of course achievement is more meaningful to both instructors and enrollees if information about the entry level skill of participants is available from pre-tests or some similar assessment procedure. Course instructors can more readily interpret the post test achievement data resulting from a course if information about the individual enrollees backgrounds has been collected. Information about prior course work in prerequisite areas, the degree to which basic concepts prerequisite to the course are used in daily work activities, and the prior formal education of participants is useful for revealing differential effectiveness of the course for different persons.

Program Administration Needs

The administrators responsible for the operation of continuing education courses also have strong interests in all of this information. In addition they need to know if the course was taught effectively in terms of the instructor's behavior, interpersonal style, and general competence both in the content of the course and in the teaching of the class. Other information about how the course was advertised, how participants heard about the course, and how adequate the location, physical setting, and time period used to teach the course are all important information to persons who operate such programs. Consequently information of this type needs to be collected, tabulated, and used to make decisions about future course offerings, instructor assignments, scheduled locations, and optimal durations. The perceptions of course participants, the persons who sent them to participate in the course, and accurate records and descriptions of the conditions under which courses operate are all important sources of information which need to be collected, tabulated, and summarized for the purpose of making these types of decisions.

Client Agency Needs

Information about the operation of a course, the characteristics of its participants, the effectiveness of the instruction in achieving intended learning outcomes, and related matters are also of interest to the agencies and companies who send engineers to participate in continuing education courses, as well as to the prospective participants. In addition, the former groups are interested in the qualifications of the course instructors, the adequacy of the instructional materials and facilities, and the committment of the continuing education unit or program to continue to work with and support the learning needs of engineers in specific regions and companies. For this reason companies seek information about the reputation of the continuing education program, the courses, and the instructors who teach them. The tacit evaluation of the worth and effectiveness of the particular course and the credibility of the sponsoring institution is important when decisions are made whether or not to involve one's employees in that course. Information about the number of replications of a course with different groups and testimonials from satisfied former individuals and their companies often provide the basis for such decisions.

Professional Societies' Needs

Professional engineering societies concerned with awarding CEUs for successful short course participation are interested in the qualifications of continuing education program sponsors, the quality of the courses taught, and the qualifications of the instructors who teach the courses (Council on the Continuing Education Unit, 1979; Martin & Greenfest, 1980). Interest in the specific learning outcomes on specific aspects of courses is not of great concern to these groups. Rather, if the creditability of the institution offering a continuing education program is established, the assumption is usually made that the instruction is valuable. It is expected that the course instructor will make a valid assessment of individual students' learning and assign some sort of grade which qualifies or does not qualify the individual student for CEU credit (Enell, 1980). This is a common pattern in higher education involving the accreditation of institutions which then offer programs and make judgments about the degree of individual student success in meeting objectives for specific courses by whatever means.

Meeting Diverse Information Needs

Meeting these diverse information needs requires thorough documentation related to planning, conducting, and evaluating continuing education courses. Information about the abilities of enrollees to perform the specific tasks being taught in a course upon entry to and departure from the course is basic. Yet by itself this information is not sufficient to explain why learning did or did not occur. The differential effectiveness of courses on the same topic, or of the same course for different persons, or of the same courses for similar groups of persons with different instructors and under different instructional conditions cannot be explained by pre- and post test or other forms of learning achievement assessments alone. Rather, information about the conditions under which instruction and learning have occurred, as well as about other variables mentioned in the earlier sections of this chapter, is needed to interpret _why_ the observed learning outcomes result. It is essential to keep good records concerning the details of when continuing education courses are offered; who is assigned to teach them; how they are advertised; what formats are selected for their presentation; how many students are enrolled; and how the content and instructional materials are developed, selected, and presented. The forms included in Appendix A are designed to collect data of this kind and serve as one set of examples of how to systematically gather information about course characteristics.

The need for large amounts of information does not require that every participant in every replication of a course be asked to complete every survey instrument, interview form, and available test. Rather as was suggested in earlier chapters, it is expedient to sample persons within courses and the employers of these persons to collect some of this information. After a course has been in operation for some time, and its properties and characteristics have been determined and found to be appropriate, it is only necessary to assess individual student learning outcomes on the course content and to monitor occasionally other aspects of participants, course instructors, and course operation, as this information is needed for reporting program characteristics and achievement to various groups.

Basic Information: Student Achievement, Course, and Instructor Characteristics

The basic means for assessing and reporting learning outcomes suggested in this book is to use the results of pretests, embedded tests, post tests, and delayed post tests. Among this array of tests, pre- and post tests are most useful for making inferences about the degree of individual student learning resulting from a course and also the general level of course effectiveness across persons. Embedded test tasks are more useful in the ongoing instructional activities of a course for instructional decision making. Delayed post tests are most useful for examining long term course effects upon participants' functional knowledge and skill in applying course content to problems in a work setting.

Procedures for developing these types of tests, and means to insure their validity and reliability, are outlined in Chapters 6 through 11. If data from such tests are collected, a variety of judgments can be made about the effectiveness of the course in reaching its intended learning outcomes. Information about the progress of individual learners can be reported to them and to persons they designate. Test data also can accumulate in various ways and be used to make improvements in the organization, content, and teaching of the course (formative evaluation). Test data can be aggregated to provide information about the overall effectiveness of the course in meeting its objectives (summative evaluation). Additional non-test data concerned with participant characteristics, instructor characteristics, and course operation can be collected with instruments similar to those presented in Appendix A. This descriptive data can be tabulated and used to explain the effectiveness or lack of effectiveness of instruction, toward improving the quality of future replications of specific courses.

In summary, pre- and post test scores of individuals enrolled in a course are the basic measures of achievement by which to judge the degree of learning of individuals and general level of course effectiveness. Additional data collected on course operating characteristics and instructor competence and behavior are not measures of learning outcomes, but are necessary to understanding the observed learning outcomes measured by the pre- and post test instruments. Without the second set of non-test data it is not possible to explain the differential effectiveness of courses for different persons, under different conditions, and with different instructors. Decisions about courses, their organization, pacing, scheduling, and staffing all require both the basic achievement data and the descriptive data which can be gathered by procedures similar to those described in Appendix A.

Examples of ways to report learning outcomes will now be presented. It should be noted that learning outcomes based upon pre- and post test or delayed post test scores are meaningful only to the degree that the tests are reliable and valid. As was pointed out in Chapter 10 this means that the course objectives being measured by the test need to be stated in performance terms; test items need to be mapped to the full range of performance objectives; a matrix of objectives by test items by topics is needed to insure that the test is representative of the learning outcomes expected to result from the course; the test tasks must be abbreviated and time efficient; and that the test tasks should be externally validated. The reliability of the test items also needs to be established. Assuming all of the above procedures have been carried out, the inferences about degree of learning resulting from a course based upon pre- and post test scores of individuals can be quite strong, especially if the results are aggregated across persons. If the results replicate from one teaching of the course to results obtained with other groups, the inference about course effectiveness in achieving desired learning outcomes can be very strong.

In the examples used, pre- and post test data for each student enrolled in the course will be the basic information used to make decisions about the degree of learning achieved by individuals and on the average by groups. Delayed post test data is generally hard to collect and is usually not collected across all individuals. Rather, delayed post tests or other assessments of performance after the course has been completed for some time are usually sampled across persons and courses for purposes of making inferences about the long term effects of courses upon performance.

Gathering and Presenting Basic Achievement Data: An Example

Let us consider a short course of about three hours duration. The course is titled "Urban Storm Water Quality Modeling: Removal and Impact." The course was one of three short courses offered during the Sixth Annual International Symposium on Urban Storm Runoff, held at the University of Kentucky in July, 1979. The course included two hours of formal instruction followed by one hour of example problem solving and discussion. The problems of evaluating courses of this short duration are somewhat special because of the limited time available. Yet, the procedures for collecting and reporting learning assessment data are basically the same as for other courses.

This particular course was evaluated by the Learning Outcomes Measurement Project as one of the activities of the group. The course was taught by Dr. Michael Meadows, a civil engineer in the College of Engineering. The purpose of the course was to impart new concepts and skills in the topic of urban storm water quality modeling. The course is an example of the third category of courses described in Chapter 2, being concerned with imparting advanced technical concepts and skills to practicing engineers, technologists, and scientists. The course had never been offered before. Participants successfully completing the course were to be awarded 0.3 CEU from the College of Engineering, University of Kentucky. A total of 31 persons completed the course in three different replications, with eight persons in the first session, fifteen in the second session, and eight in the final session. The participants were typically civil engineers working as consultants or for federal, state, or local government agencies.

The major intention of the evaluation was to determine the effectiveness of the course, rather than to make strong individual assessments of each student's learning. The reason for this is that the time available for instruction was very short. No more than a few minutes could be devoted to testing with a maximum of ten minutes allocated for the pre-test and another ten for the post test.

The course instructor developed twelve test items of the essay or constructed response type. The twelve items were sorted into three categories: easy items, moderately difficult items, and difficult items. Test items not meeting these criteria were rewritten in order than an equal number of items in each category was obtained for each test form.

One test item from each category was randomly assigned to an individual form of the test. Four forms were constructed for use as pre-tests and post tests. The actual test questions and their assignment to the four different forms of the test are shown in Table 7.

An item sampling procedure was carried out to insure an adequate assessment of the learning outcomes of the course. This is consistent with procedures designed to estimate the effects of courses on achievement of persons generally (Lord & Novick, 1968; Shoemaker, 1973). However, the performance of each individual on these tests is only a partial estimate of each individual student's learning in relation to the total course content. Yet, since several individuals responded to each question on the pre-test and several other items on the post test, a relatively accurate estimate of the entry knowledge level of participants as a group and the growth in the knowledge following the course activities can be obtained.

The course instructor graded each question in terms of the knowledge displayed in the answer which had been constructed by the individual participant. Three scoring categories were used, with a 0.0 value assigned for answers

Table 7

Pre- and Post Test Items and Their Assignment to Test Forms for the Urban Storm Water Course

1. How are the parameters for stormwater pollutant washoff models determined?
2. Which stormwater pollutant washoff model are you familiar with? What is your major criticism of this model?
3. What is the state-of-the-art model (equations) for routing unsteady streamflow?
4. In stormwater and quality modeling, what does the term "regionalization" mean?
5. When are the celerities of streamflow and water quality models the same?
6. Distinguish between dynamic and kinematic waves.
7. How can a person best simulate the dynamic response of a receiving stream's water quality system to stormwater pollution?
8. Are all one dimensional streamflow and water quality routing models compatible? Explain your answer.
9. What model would you recommend for routing streamflow during periods of stormwater runoff?
10. What process(es) is (are) involved in the washoff of pollutants from an urban watershed?
11. How can data collected at one watershed be transferred to another watershed?
12. During the preliminary assessment phase of an area-wide quality study, how can a person identify those land use areas that are potentially significant sources of stormwater pollution?

FORMS	PRE-TEST QUESTIONS	POST-TEST QUESTIONS
A	1, 2, 3	3, 9, 11
B	4, 5, 6	2, 4, 12
C	7, 8, 9	5, 7, 10
D	10, 11, 12	1, 6, 8

with no knowledge of the concept, 0.5 being used for scoring the presence of some knowledge, and 1.0 being assigned for correct knowledge of the question content. Each item was scored in this manner for each person. The total possible score for any person was 3.0 and the minimum score possible was 0.0. Responses to both the pre- and the post test were scored in this manner by the instructor and the results for individual students on the four test forms were recorded.

The equivalence of the various forms of the test in the pre- and post test roles was determined by a one-way analysis of variance across the scores of participants on the four forms of the test. In this procedure it is assumed that because of random assignment of participants to test forms on both the pre- and post test, all participant groups can be viewed as being equivalent in terms of their expected mean scores on the tests. Therefore, if there are statistically significant differences between the four forms of the test in either the pre- or post test role, the non-equivalence of the test forms would be suggested. In light of such findings, the test items and forms might need to be reworked.

Table 8 contains the observed total scores for the 33 persons who began the course and were randomly assigned to one of the four forms in the pre-test role. Persons from all three replications of the course were pooled for the analysis. In Table 8 the actual test scores for persons across the four

Table 8

Pre-Test Total Scores Across Test Forms
Urban Storm Water Quality Modeling Course

	Test Form			
	A	B	C	D
	3.0	1.0	1.5	1.0
	0.0	0.0	0.0	2.0
	3.0	0.0	0.0	2.5
	1.5	0.0	0.0	2.5
	0.5	0.0	0.5	1.0
	1.0	0.0	2.0	2.0
		2.0	0.5	
		0.0	0.5	
		0.0	1.0	
		0.0		
		0.5		
		1.0		
Statistic				
n	6	12	9	6
$\bar{X}$	1.500	0.417	0.667	1.833
s	1.265	0.634	0.707	0.683

One Way Anova Table

Source	SS	df	MS	F	p≤
Between Forms	10.629	3	3.543	5.480	0.004
Within Forms	18.750	29	0.647		
Total	29.379				

forms of the pre-test are listed along with the mean scores and standard deviations for each form. Following this information a one-way analysis of variance table is presented. It is apparent that the four forms of the test are not equivalent for the 33 persons who completed the pre-test. Forms D and A were the easiest and forms B and C the most difficult. Of course, these findings could be due to differences in the entry level knowledge of participants in the four groups even though they were randomly assigned. By chance alone, more able persons may have been assigned to forms A and D. However, the test items and forms should be reworked to insure more equality of assessment of entry level knowledge in the course content, since the statistical test shows that the probability of such a distribution of scores under random assignment of persons to test forms is unlikely.

Table 9 is a similar presentation of the individual total scores of the 31 participants who actually completed the course across the four forms used as a post test. Again the individual scores of persons and test forms, standard deviations, and an analysis of variance table are presented. Persons from all three replications of the course are pooled. The results indicate that in the post test role the four forms of the test are not statistically significant from one another. Of course, this may be caused by the fact that the instruction in the course has caused all persons to master

Table 9

Post Test Total Scores Across Test Forms
Urban Storm Water Quality Modeling Course

Test Form

A	B	C	D
2.0	2.5	2.5	3.0
2.5	2.0	2.0	3.0
3.0	3.0	2.0	2.0
3.0	2.5	2.5	1.0
2.0	2.0	1.5	3.0
3.0	3.0	2.5	2.5
2.5		1.5	2.5
		2.0	2.0
		3.0	1.5

Statistic	A	B	C	D
n	7	6	9	9
$\bar{X}$	2.571	2.500	2.167	2.278
S	0.450	0.447	0.500	0.712

One Way Anova Table

Source	SS	df	MS	F	$p \leq$
Between Forms	0.827	3	0.276	0.900	0.454
Within Forms	8.270	27	0.306		
Total	9.097				

the material to such a high level that differences in the difficulty level of the four forms of the test no longer exist. Or it may be that the four forms of the test are equivalent.

For purposes of this example, it is assumed that the four forms of the test are equivalent. It is also assumed that the tests have been properly constructed, and that they are both valid and reliable. Administration of the tests in the pre- and post test roles to other groups and subsequent analyses would help confirm or refute these assumptions, as would the carrying out of appropriate item analysis and test form reliability and validity studies (See Chapters 10 and 11). It should be noted that the method of assigning persons to test forms in both the pre- and post testing, although random, prevented any person from taking the same form of the test in both the pre- and post test situation.

Figures 2, 3, and 4 show the results of the three replications of the course on both pre- and post tests. Pre- and post tests scores are plotted against the rank of persons in each section of the course. Persons in Groups 1 and 3 have been ranked in order of the overall quality of their total written responses on the post test, a procedure carried out in addition to the individual grading of items. Since there are 8 persons in each group, there are 8 ranking categories. For Group 2, which had 15 persons, the actual observed post test score categories are used for the ranking.

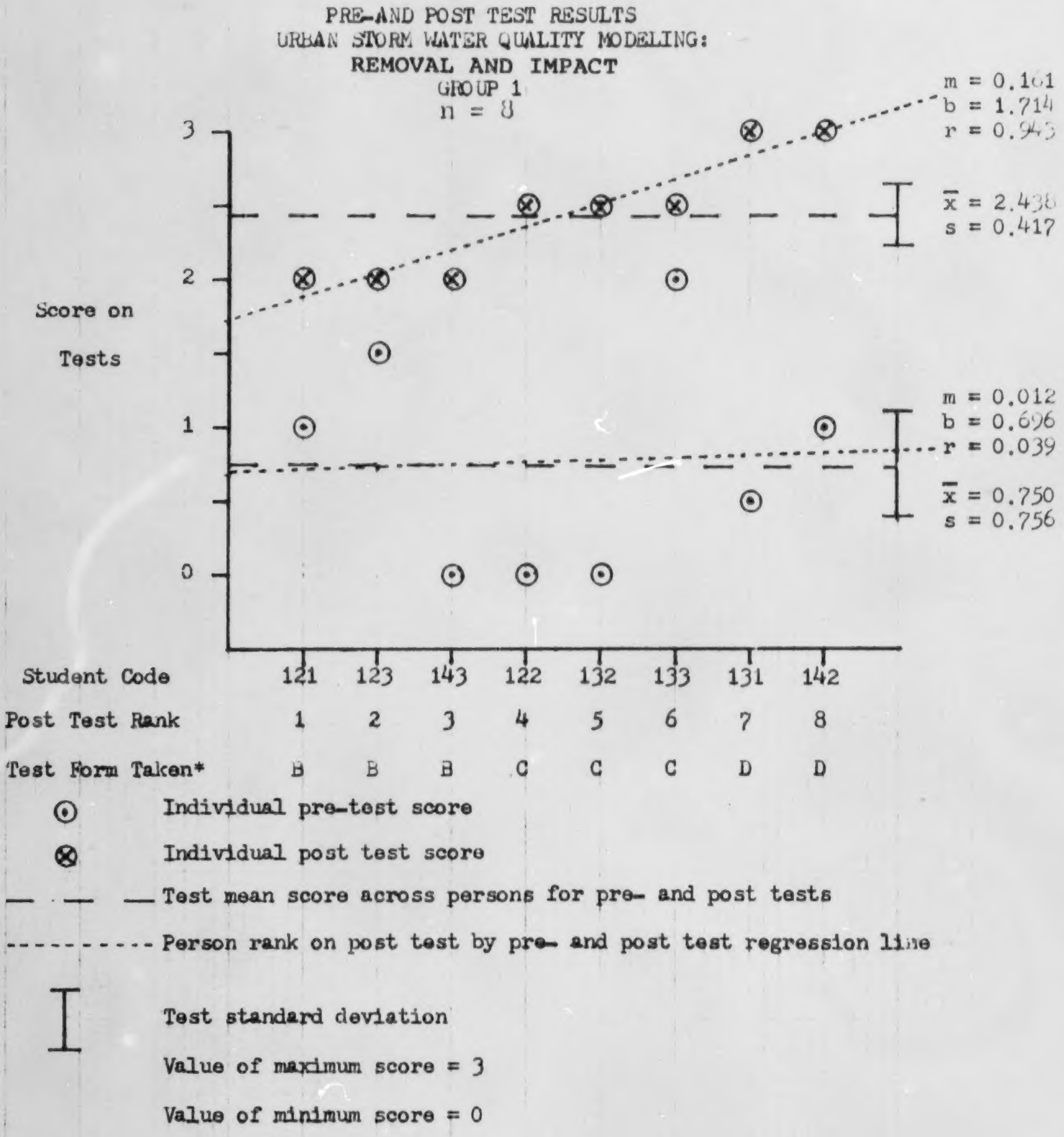
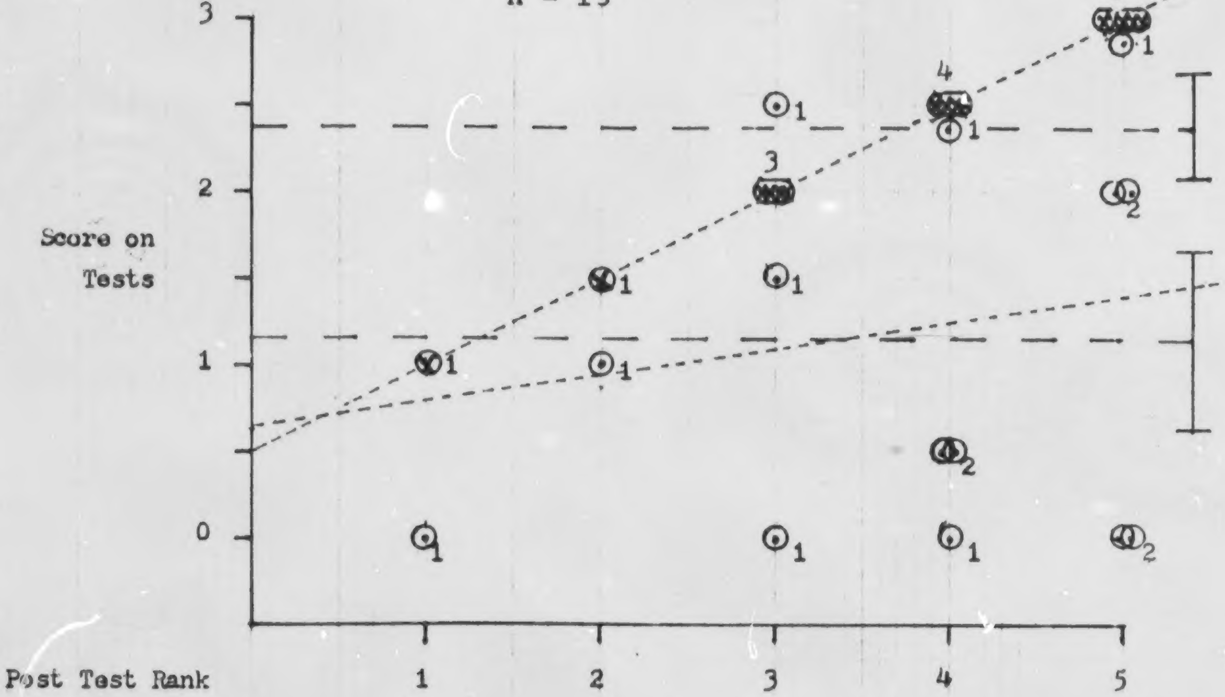


Figure 2 Illustration of a graphic means for reporting learning outcome measurements for a short course to individual participants and groups.

*See Table 7 for determination of test form

PRE- AND POST TEST RESULTS
URBAN STORM WATER QUALITY MODELING:
REMOVAL AND IMPACT

GROUP 2
n = 15



Individual pre-test scores

Individual post test scores

Test mean scores across persons for pre- and post tests

Person rank on post test by pre- and post test regression lines

I Test standard deviation

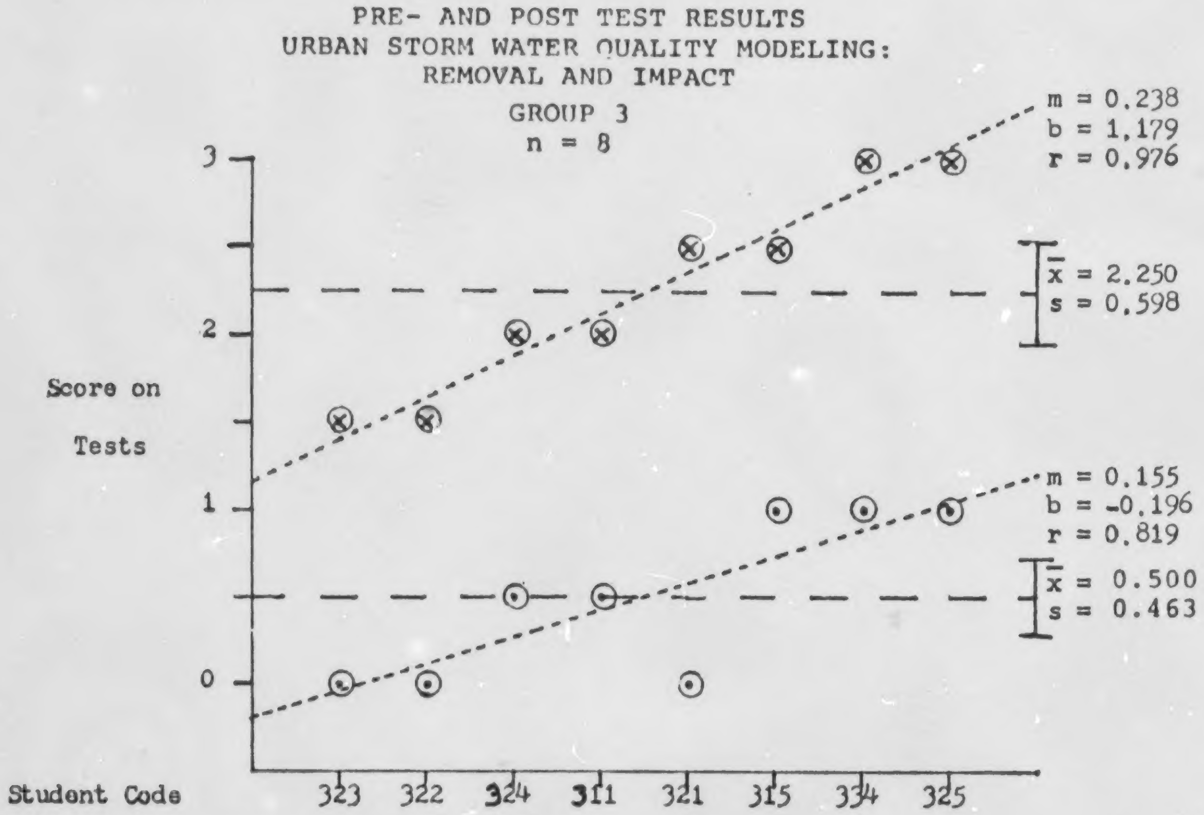
Value of maximum score = 3

Value of minimum score = 0

	Pre-Test	Post Test
m	0.195	0.500
b	0.513	0.500
r	0.195	1.000
$\bar{x}$	1.167	2.367
s	1.097	0.611

Figure 3 Learning outcomes resulting from a short course for engineers on Urban Storm Water Quality Modeling.

* Individual persons pre- and post test scores across forms cannot be listed in this figure because persons are ranked by score categories rather than by individual persons.



Student Code	323	322	324	311	321	315	334	325
Post Test Rank	1	2	3	4	5	6	7	8
Test Form Taken*	A	A	B	B	B	B	B	C

Individual pre-test score

Individual post test score

Test mean score across persons for pre- and post tests

Person rank on post test by pre- and post test regression line

I Test Standard deviation

Value of maximum score = 3

Value of minimum score = 3

Figure 4 Learning outcomes resulting from a short course for engineers on Urban Storm Water Quality Modeling.

* See Table 7 for determination of test form.

Since only 5 of the 7 possible total score categories occurred in Group 2, there are 5 ranks.

A mean, standard deviation, and regression line are presented for both the pre-test and the post test for each of the three groups. The information is presented graphically as well as numerically. A quick glance at the three figures reveals the basic patterns. The post test means are seen to be uniformly high and about the same value for all three groups. The pre-test means are seen to be much lower and about the same value for Groups 1 and 3, but to be higher for Group 2. The greater variability of the four forms of the test across the persons on the pre-test is immediately apparent, especially for Groups 1 and 2. The most striking feature of the graphs is the consistent and large difference between the pre-test and post test means across the three groups.

Data of this type collected on well designed pre- and post tests across replications of courses is strong evidence of the degree of learning which has resulted from the course. Presented in graphic form it is much more interpretable than if simply presented numerically. Commercial computer programs exist which allow for the easy tabulation of pre- and post test data and the construction of plots similar to those presented in Figures 1 through 3. The information gained from such data is useful for both formative and summative evaluation procedures. It can be used for reporting results of learning outcomes to many groups of persons, including course participants, course instructors, program administrators, client agencies who sent participants, and accrediting groups from professional organizations. It is the most basic data which can be obtained about the learning outcomes of the course. Without this or similar information about the performance capabilities of course participants at the beginning and end of the course, little can be said with assurance about the degree of learning which has resulted for individuals or for groups taken as a whole.

For these reasons the evaluation of individual learning outcomes or the evaluation of course effectiveness generally needs to be based upon some similar procedure. The data collected is not only useful to reporting the individual achievements of particular students to them, but collected over replications of a course and over many courses within a program, it can be very effective in making assessments of the effectiveness of courses, various instructional organizations and arrangements, and the credibility of programs of continuing education offered by universities or other agencies.

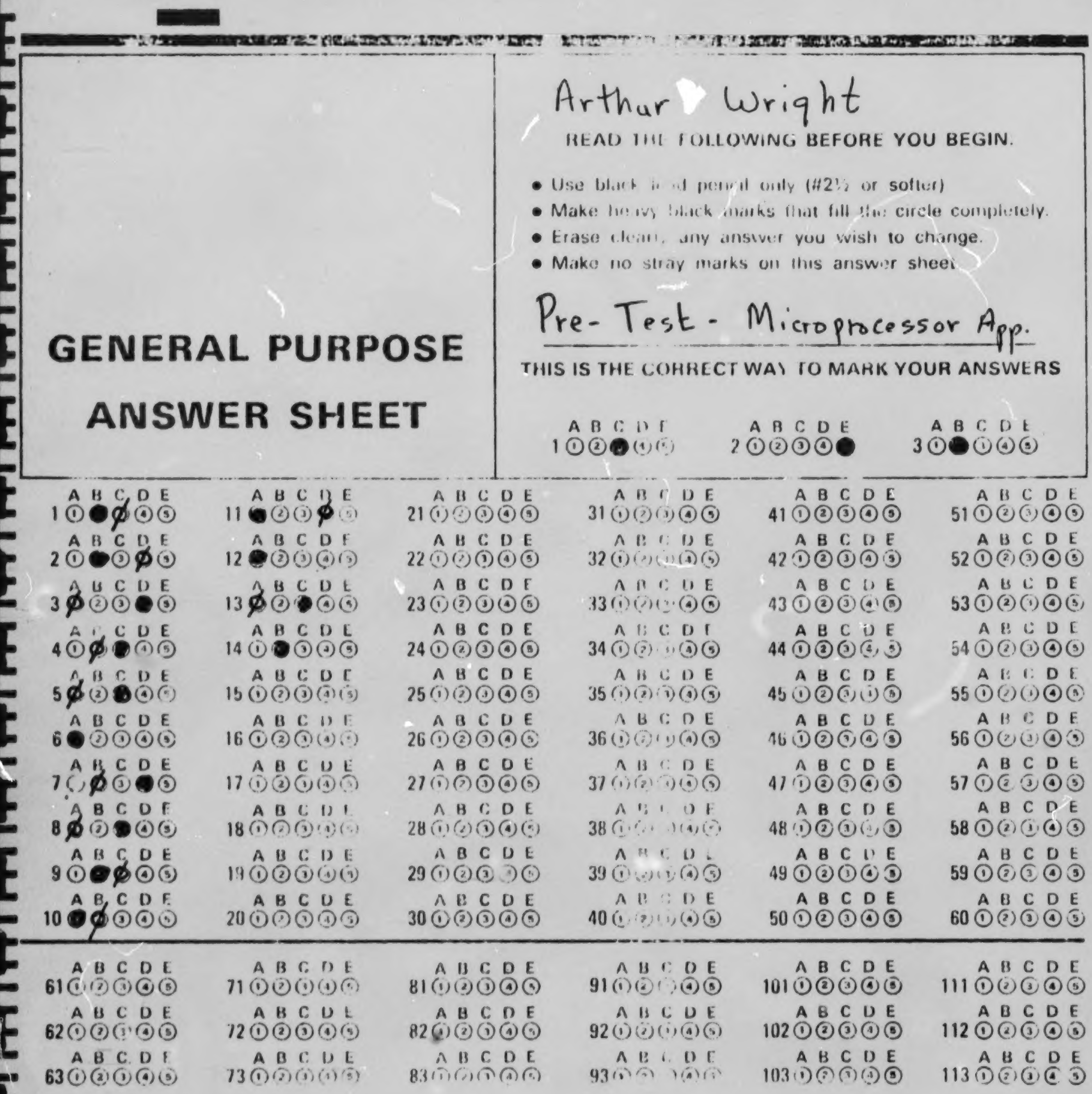
Reporting Learning Outcomes to Individual Students

It is well established that immediate feedback concerning the accuracy and adequacy of performance facilitates student learning and motivation. Course instructors should always provide students with the results of corrected homework problems, quizzes, laboratory exercises, and other types of embedded test tasks as soon as possible after they have been administered. Immediate knowledge of results is desirable in these situations. This procedure allows individual students to compare their own recently completed performances to those presented by the instructor. Oftentimes the instructor cannot provide students immediately with their corrected and scored homework problems, exercises, and quizzes because time is needed to complete the correction of student responses. An alternative procedure is to hand out common sets of correctly worked problems, solutions to test problems, and laboratory exercises. This method achieves immediate feedback to students about the adequacy of their recently completed performance against a detailed example of how the problems or exercises should have been completed. However, correcting and grading of individual students' homework and other papers and prompt return of these to students remains important.

In short courses, similar to those often used in continuing education activities in engineering, time is so limited that it is difficult to correct promptly students' work and to arrange adequate opportunity for individuals to examine and reflect on these results. Some of the methods for overcoming this obstacle are noted in Chapter 7 in the discussion of embedded testing procedures. It is important that both students and instructors quickly obtain knowledge of results from embedded test tasks in short courses. Otherwise the assessment procedures serve no useful function for guiding learning and instruction. The Haan and Barfield (1978) "Hydrology and Sedimentology of Surface Mined Lands" course, as it is described in Chapter 7, is one good example of how to provide students in short courses with immediate knowledge of the results of their performance. This course makes use of embedded test tasks, sample problems, and completely worked solutions handed out as soon as participants have completed assigned work.

One advantage of the multiple choice test format or other short answer objective tests is that they can be scored immediately after students have completed them. A standard scoring sheet can be used by students to mark the appropriate answer to each question. The answer sheet can be scored by machine immediately, right in the classroom, if the proper equipment is available. Equipment for this purpose is currently commercially available. However, even without such equipment, standardized answer sheets for multiple choice questions can be scored by hand using a scoring over-lay or a master answer sheet. A test with as many as thirty items can be scored in as little as 15 seconds by this procedure. Furthermore, the correct answer to each question can be marked on the student's paper by marking through the opening on the master or answer sheet. The total number of errors can be counted as scoring proceeds and be noted on the student's answer sheet and recorded by the instructor. The scored answer sheet with the correct responses to items the student missed can be returned to the individual student a few seconds after he or she has completed the test. The student can be allowed to go over his or her own test using the corrected answer sheet and the test booklet which contains the questions and problems. In addition, a solution sheet can be provided which explains why a particular answer is correct to each question on the test and why the distracting options for each question are wrong. Use of this procedure is extremely effective in providing students with information about their performance. Tests carried out in this manner are very instructive to students who quickly identify what specific areas or concepts they do not yet understand and need to learn to master. The immediate scoring and recording of the test results by the instructor also alert him or her to problems that individuals or groups are having. Often, immediate corrective action can be taken by the instructor when the testing is completed in order to remedy problem areas.

Figure 5 is an example of an actual answer sheet scored by a hand key and returned to the student immediately after the test was completed. The scoring key is simply another answer sheet with spaces punched out for the correct answer



Arthur Wright

READ THE FOLLOWING BEFORE YOU BEGIN.

- Use black lead pencil only (#2½ or softer)
- Make heavy black marks that fill the circle completely.
- Erase clean, any answer you wish to change.
- Make no stray marks on this answer sheet.

Pre-Test - Microprocessor App.

THIS IS THE CORRECT WAY TO MARK YOUR ANSWERS

A B C D E

1 ①②●④⑤ 2 ①②③④● 3 ①●③④⑤

GENERAL PURPOSE ANSWER SHEET

1 ①●ø④⑤	11 ●②③ø⑤	21 ①②③④⑤	31 ①②③④⑤	41 ①②③④⑤	51 ①②③④⑤
2 ①●③ø⑤	12 ●②③④⑤	22 ①②③④⑤	32 ①②③④⑤	42 ①②③④⑤	52 ①②③④⑤
3 ø②③●⑤	13 ø②●④⑤	23 ①②③④⑤	33 ①②③④⑤	43 ①②③④⑤	53 ①②③④⑤
4 ①ø●④⑤	14 ①●③④⑤	24 ①②③④⑤	34 ①②③④⑤	44 ①②③④⑤	54 ①②③④⑤
5 ø②●④⑤	15 ①②③④⑤	25 ①②③④⑤	35 ①②③④⑤	45 ①②③④⑤	55 ①②③④⑤
6 ●②③④⑤	16 ①②③④⑤	26 ①②③④⑤	36 ①②③④⑤	46 ①②③④⑤	56 ①②③④⑤
7 ①ø③●⑤	17 ①②③④⑤	27 ①②③④⑤	37 ①②③④⑤	47 ①②③④⑤	57 ①②③④⑤
8 ø②●④⑤	18 ①②③④⑤	28 ①②③④⑤	38 ①②③④⑤	48 ①②③④⑤	58 ①②③④⑤
9 ①●ø④⑤	19 ①②③④⑤	29 ①②③④⑤	39 ①②③④⑤	49 ①②③④⑤	59 ①②③④⑤
10 ●ø③④⑤	20 ①②③④⑤	30 ①②③④⑤	40 ①②③④⑤	50 ①②③④⑤	60 ①②③④⑤
61 ①②③④⑤	71 ①②③④⑤	81 ①②③④⑤	91 ①②③④⑤	101 ①②③④⑤	111 ①②③④⑤
62 ①②③④⑤	72 ①②③④⑤	82 ①②③④⑤	92 ①②③④⑤	102 ①②③④⑤	112 ①②③④⑤
63 ①②③④⑤	73 ①②③④⑤	83 ①②③④⑤	93 ①②③④⑤	103 ①②③④⑤	113 ①②③④⑤

Figure 5 Sample Standard Answer Sheet for Manual or Machine Scoring

● Indicates the student's individual response to each item

ø Indicates the correct response as scored by the instructor using the scoring key, when the student has responded incorrectly.

to each question. If no standard answer sheets are available, suitable answer sheets can be constructed by typing rows and columns of zeros or capital "Os" on a plain sheet of paper and adding numbers for rows and letters for options within rows. The master or scoring key is made and used in the same manner as is the case with the sample answer sheet in Figure 5.

Similar procedures can be used with other objectively scored test items. If the test items result in a particular numerical value, the construction of a particular diagram, or in some other standard response which can be quickly and objectively scored, relative immediate knowledge of results can be communicated to students by scoring tests and returning them to students as soon as they are completed.

It should also be recalled that multiple choice test items can be used for any type of testing situation. The sample test in Appendix B illustrates how complex performance capabilities may be tested for by well designed and abbreviated test tasks. Much is said about this in earlier chapters. It will suffice to note that if more attempts were made to cogently encapsulate the basic features of the intended learning outcomes of courses within well constructed test items similar to those shown in Appendix B, it would be much easier to score student performance immediately after assessment and communicate the results of individual's performance to them at once.

At the end of courses it is also important to communicate the achievement results to the individual. Information about how much each person had learned with respect to his entry level knowledge is of interest to the participant. If the course is small and the pre- and post test scores of individuals have been plotted against some criterion of performance capability (e.g., Figures 2 and 4) this information can be shared with individual students following the course. In both Figures 2 and 4 each student can identify him or herself by the student code number. The pre- and post test scores of the individual can be identified. The person's performance in relation to other persons in the group and in relation to the degree of mastery can all be determined.

For courses in which it is not possible to list each individual's performance on a group performance graph, it is still possible and important to provide participants with information on their own performance in the course. Figure 6 is an example of a standard form which can be used to report the achievement of individual students in courses to persons at the completion of a course. Typically it might require a few days to compile and prepare all of the individual achievement reports for a course. These can be mailed out to students at the conclusion of the course. With efficient scoring and processing procedures, it is often possible to provide participants with this type of information prior to their departure from a course. In any event, the

Figure 6

MANUAL INDIVIDUAL ACHIEVEMENT REPORTING FORM

Name: Brown, Charles P. Student Code #122

Course: Urban Storm Water Quality Modeling: Removal and Impact

Instructor(s): Dr. Michael Meadows

Date(s): July 24, 1979

Your performance on the pre-test and post test is reported below in both graphic and numerical form. In addition, information on the performance of the class as a whole is also recorded.

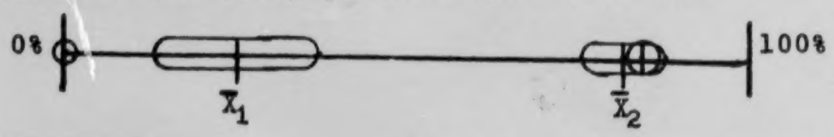
Class Performance

Pre-test (x_1)		Post test (x_2)	
n =	8	n =	8
$\bar{x}$ (%) =	25.00	$\bar{x}$ (%) =	81.27
s.d. (%) =	25.20	s.d. (%) =	13.90

Your Performance

	Pre-test (x_1)	Post test (x_2)
% Score	0.00	83.33
Percentile Rank	37.50	62.50

Plot of individual and group performance: Pre- and post test



n = no. of participants
$\bar{x}$ = mean test score
s.d. = standard deviation ⬭
⊕ Your pre-test score
⊕ Your post test score

We hope this information will be beneficial to you. Should you have questions or comments please contact this office.

---

*Percentile ranks based upon rank order scores of pre- and post test scores shown in Figure 2, page 278.

opportunity to share with individual students the results of their performance on the pre-test and the post test should not be dependent upon completion of the achievement report. Rather, by using the methods described above, these tests can also be scored immediately and the results communicated to the students with corrected answer sheets, the test question booklets, and solution sheets. If the test results are not shared with students immediately in this or some similar manner, the instructional value of the testing procedure is lost.

Even when students have received immediate feedback on their performance on learning assessment activities, there is still a need to make a summary report to each student at the end of the course. Figure 6 is a completed achievement report for one student from the Urban Storm Water Quality Modeling Course. The student may be identified in Figure 2. The name listed in Figure 6 is ficticious, but the results are for a real person enrolled in this course. The form is designed for easy use by clerical staff involved with the continuing education program. Only the information which is written in by hand needs to be prepared for each individual student. The remainder of the information pertains to the learning outcomes for the entire course. If a course has many participants enrolled, the information which applies to the entire course can be typed on a master copy for that course.

Sufficient copies can then be duplicated to prepare an individual report for each enrollee. The specific information for each individual can quickly be added with minimum effort. This is illustrated in Figure 6 by the handwritten information.

The name of the course, the instructor and the data can all be typed in on a master. The same procedure can be followed for the information about the class performance. In addition, the pre- and post test means and standard deviations can also be plotted on the master. All the copies needed for the total enrollment of the course can then be duplicated. What remains is to simply add the information needed for any individual. This amounts to adding the person's name and student code number, listing the person's pre- and post test scores and percentile ranks, and plotting the person's pre- and post test scores on the line which already contains the group means and standard deviations.

The line at the bottom of the form on which the group pre- and post test means and standard deviations are plotted along with the individual's scores on the two tests is facilitated by the length of the line and the metric in which the test scores are reported. The line is 10 centimeters long. The scoring metric is in percentage of the total possible score. It is a simple matter to plot an individual's score directly as a percentage on the 10 centimeter scale with a metric ruler. If time is short, the group means and standard deviations can be plotted on the master form for the group, and the individual be instructed in a standard comment on the form to plot his or her own scores on the line if so desired.

The form is simple to use. It provides individual participants with the basic information they need concerning formal assessment of their learning outcomes in a course. In addition, if a form similar to this one was prepared for each person in each course in a continuing education program over many offerings of courses, much information about course effectiveness over replications and much information about program effectiveness across courses could be accumulated. Although such information alone will not serve to replace the tacit evaluation of courses and the institutions which offer them by the client agencies and groups who enroll in continuing education courses, such information can be very useful in a supporting way. It can also serve as an aid to quality control, improvement of courses and programs of study, and evidence of the worth of continuing education activities operated by universities, colleges and other units. Accrediting agencies, professional societies, and governing boards and groups all have legitimate interests in this type of achievement data aggregated over courses.

The task of preparing the individual achievement report can be simplified by computer processing. Figure 7 is an example of a computer prepared report. The report is for a student enrolled in one section of the "Hydrology and Sedimentology of Surface Mined Lands". In this case, the

Figure 7

Computerized Individual Achievement Reporting Form

```
*************************************
*                                   *
*      UNIVERSITY OF KENTUCKY       *
*      COLLEGE OF ENGINEERING       *
*  OFFICE OF CONTINUING EDUCATION   *
*     LEXINGTON, KENTUCKY 40506     *
*       (TEL: 606-257-3971)         *
*                                   *
*************************************

COURSE NAME:      HYDROLOGY AND SEDIMENTOLOGY FROM SURFACE MINED LANDS
SITE OR SECTION:  TULSA, OKLAHOMA
DATE OF COURSE:   JUNE 19-21, 1979
INSTRUCTOR(S):    DR. BILLY BARFIELD AND DR. TOM HAAN

PARTICIPANT:      ______________________________________________

**********************************************************************

          _______________CLASS PERFORMANCE_____(N = 30)___

                                 _PRETEST_  _POSTTEST_  _GAIN_

          AVERAGE TEST SCORE       10.8        16.0       5.2
          STANDARD DEVIATION        3.1         2.9       3.4
          TEST RELIABILITY               .685
          STD.ERROR OF MEAS.              1.7
          PERCENT OF THE CLASS      6.0        60.0
            MASTERING THE TEST
          ________________________________________________________

          YOUR TEST SCORE           13          18         5

AS PART OF THE ABOVE COURSE, YOU WERE GIVEN TWO 22 QUESTION MULTIPLE
CHOICE TESTS.  THE TEST QUESTIONS WERE CLOSELY RELATED TO THE MATERIAL
COVERED DURING THE COURSE.  THE CLASS WAS NOT EXPECTED TO DO WELL ON
THE FIRST OR PRE-TEST. (IF IT WERE OTHERWISE, THE COURSE MIGHT
HAVE LITTLE VALUE TO PARTICIPANTS.) ON THE OTHER HAND, THE CLASS WAS
EXPECTED TO DO QUITE WELL ON THE POSTTEST WHICH WAS ADMINISTERED AT
THE END OF INSTRUCTION.  FROM THE VIEWPOINTS OF THE INSTRUCTORS AND
COURSE DESIGNERS, A POSTTEST SCORE OF 17 OR BETTER IS EVIDENCE THAT THE
CLASS AS A WHOLE AND PARTICIPANTS INDIVIDUALLY MET THE STATED LEARNING
OBJECTIVES FOR THE COURSE.

AN INDIVIDUAL'S POSTTEST SCORE, HOWEVER, DEPENDS ON MANY FACTORS -
SUCH AS, FOR INSTANCE, ONE'S PRIOR KNOWLEDGE AND FAMILIARITY WITH
THE GENERAL DISCIPLINARY AREA AS WELL AS THE SPECIFIC SUBJECT MATTER
OF THE COURSE AS THESE ARE MEASURED BY THE PRETEST.  WHETHER OR NOT
A PARTICIPANT'S POSTTEST SCORE EQUALLED OR EXCEEDED 17 IS LESS IMPORTANT
THAN IS THE DIFFERENCE OR GAIN SCORE BETWEEN THE POSTTEST AND THE
PRETEST.  FOR THE CLASS AS A WHOLE, THIS DIFFERENTIAL GAIN MEASURES
LEARNING OCCURRING AS A BENEFIT OF INSTRUCTION AND COURSE PARTICIPATION.
FOUR OR FIVE POINTS DIFFERENCE BETWEEN PRETEST AND POSTTEST INDICATES
SIGNIFICANT IMPROVEMENT AS A RESULT OF ATTENDING THE COURSE.

WE HOPE THIS INFORMATION WILL BE BENEFICIAL TO YOU. SHOULD YOU
HAVE QUESTIONS OR COMMENTS, PLEASE CONTACT THIS OFFICE.
```

report is for a real person enrolled in the class. In the interest of privacy, the person's name and address, which would normally be listed on the computer printout, is omitted. All the basic information is given about the individual's performance and the performance of the group. In addition, test reliability and standard error of estimate values are given. For classes with large enrollments the form could be prepared directly from data students generated in their response to multiple choice questions on standard and machine scorable answer sheets. There are many ways to automate the processing of test data and the preparation of individual achievement reports.

Keeping Learning Outcomes of Individuals Private

While individuals enrolled in courses should have the results of their performance on assessment instruments and of their performance in the course as a whole communicated to them, this information is not properly communicated to anyone else (Tyler & Wolf, 1974). Participants in a given course may be sponsored by their employers. In this case, with the permission of the student beforehand, it is proper to release the individual's performance record in the course to the employer. Generally the employer will be interested in a global assessment of the individual student. In short, this translates into reporting whether the student completed the course successfully or not.

Another group to which a participant might direct that the results of his or her performance record for a course be sent are professional organizations and groups which credit and record continuing education units. Here, again, the persons who supervise such activities are generally interested in some global judgment of the course instructor's assessment of the student's performance in the course. This usually translates into some judgment about the overall adequacy of the individual student's performance in the course, either as acceptable or unacceptable for the CEU credit.

There are two points to be made. The first is that only those persons designated by individual course participants should have any information sent to them about individual student performance in the course. The second is that the information needed by these groups is of the "pass" or "fail" or "successfully completed" or "not successfully completed" type. It is inappropriate to send the individual's detailed learning report form with all of the information about the person's pre- and post test results to these other groups. The pass/fail judgment is sufficient. If the individual wishes to share the detailed report with his employer or with a professional licensing agency, he or she may do so.

This is not to say that information of the type contained in Figures 2 through 4 should not be shared with employers and persons responsible for supervision and recording of CEUs. There is no problem as long as the data is group data and what is reported is the effectiveness of the course generally for groups of students. What is inappropriate is to identify the performance of individual students in such presentations, except for use by the individual student and course instructors.

If a course has been shown to be generally effective and the persons who administer and teach it to be competent and responsible, there is no need for either employees or others to have more information about an individual student's performance in a course than to know if it was successful or unsuccessful. There is too much opportunity for misuse of detailed achievement data by employers or others if it is provided. For example, a supervisor could conceivably decide to promote one individual and not another on the basis of detailed test scores in a course and the relative rankings of the two persons. If both persons had passed the course this would be inappropriate, and it might also be inappropriate even if one person had not passed the course. There is simply too much error in individual performance test scores, even under the most ideal conditions, to make such inferences and to be correct most of the time. Other information about the individuals' skills in the work setting on a range of tasks, about their attitudes and interests, and past performance are much more crucial in making such decisions (McClelland, 1973; Stice, 1979). There are many persons who do not understand these points and who are prone to use a test score as concrete evidence for a decision which should be made by a more informed process requiring much more effort. Test scores of persons are often abused in these matters (McClelland, 1973).

The main value of test scores is their aggregation for persons over replications of courses. Used in this way, scores obtained from well designed tests can be very useful in evaluating the effectiveness of courses (Tyler, 1974). Another major use of tests is as part of the learning activities which constitute the instruction for the courses in which they are used. It is for instructional purposes that individuals' test scores should be shared with them immediately after the completion of tests and in reports to them upon course completion. The individual who has recently completed a course can interpret the results of pre- and post tests in the context of the course activities in which he or she has engaged. The scores of persons are balanced and meaningful in this context and in relation to how much the individual feels he or she has learned in areas not measured by the tests.

It is also appropriate for individuals registering for continuing education courses to decide if they choose to be involved in the testing at the pre- and post test stages. Being involved in the course will often require completing the embedded test tasks and should routinely be required as are the other activities designed to teach individuals the content and skills of the course. If persons are seeking CEU credits, and if appropriate pre- and post tests are available, these ought also to be required. If persons wish to enroll without seeking CEU credits and they wish not to participate in testing, they should be allowed to do so. About the only exception to this situation is where failure to learn to criterion some particular skill or content would result in property damage or threat to health and life. In these cases, testing should be required for all participants. An example cited in an earlier chapter is the proper assessment of individuals' competence in operating dangerous and expensive laboratory or industrial equipment before allowing them to do so certifying that they are competent to do so.

Precautions to Prevent the Abuse of Test Scores

Attention to the matters discussed in the previous section help prevent abuse of test scores. There are other precautions which should be observed to insure proper use of test scores and other performance assessments of students in continuing education courses.

Before decisions are made about the effectiveness of courses in reaching their desired objectives, the tests or or other assessment procedures used must be determined to be reasonably valid and reliable. Methods for doing so have been outlined in some detail in previous chapters. Poorly designed tests are worse than no tests. Their use may alienate students who are quick to see the invalidity of tests and test items, especially at the post test stage when they can judge how closely and how adequately the test items and assessment procedures match the content of the course.

Instructors can be helped by good tests which are appropriately comprehensive while at the same time being brief and time efficient. Imposing a poorly designed test into a short and already very full time period which is needed for instruction is a serious aggravation to instructors as well as to students. Any type of testing or assessment which is developed for a course needs to be developed with the full participation of the course instructors and should be specifically related to the course material. If such cooperation cannot be obtained there is little point in imposing an external testing or assesment procedure on a course or an instructor. Both students and instructor are apt to resent the intrusion, the results of the testing are likely to be invalid, and the data not particularly helpful to the revision of the course and its ultimate improvement. On the other hand, if the course instructor can be convinced that well designed tests and assessment procedures can be useful to promoting instruction, and can be encouraged to become involved in designing appropriate procedures, much will have been gained.

If a course is to be developed and offered many times over a period of months or years, proper testing and assessment of learning outcomes is important to the formative evaluation of the course and its quality control throughout its life time. In such cases, the investment of the initial time and effort needed to develop good tests for the course can be repaid many times over. The benefits derive from having good information concerning the operation of the course at various times, under various conditions, and with different instructors. Often much is learned in the design and evaluation of one course which is useful in the design and operation of other courses.

Sometimes, if a course is to offered only once or twice for a special group of persons, there is no need to engage in elaborate test development activities. However, in most courses instructors will have notions of what it is they expect students to learn and be able to do at the end of the course. Beginning with these expected learning outcomes is basic to the very design of any course of instruction. It is often not difficult to translate these expectations into some sort of formal assessment tasks. The Urban Storm Water Quality Modeling course presented as an example earlier in this chapter is just such a case. The course was not offered many times. It was very short, being only three hours long. The instructor had a clear notion of what he expected students to achieve as desired learning outcomes. It was relatively easy for him to put together 12 test items which would give some indication of the entry level and exit level knowledge of his students. Furthermore, the information collected was useful to the instruction of the students. In addition, it would provide better evidence of the effectiveness of the course in reaching the intended objectives than more casual information collected in a less systematic way. Therefore, it was a good idea to develop and use the pre- and post tests. The results shown in Figures 2 through 4 are certainly informative and helpful to making decisions about course effectiveness. For this course, and for similar short courses which are not to be replicated many times, it does not make sense to devote great efforts to the development of testing and assessment procedures. What was done for the Urban Storm Water course was very adequate.

The manner in which the test items were developed was also consistent with the optimum procedures described in Chapter 10, although some individual steps were omitted and the whole procedure took only a short time. To the extent that an instructor has a good grasp of what he or she expects to achieve in the teaching of a course, and to the extent that he or she is well organized in his or her instructional plans and activities, it is not difficult to sample appropriate tasks from within the course activities to be used as test items. Although the procedure looks formidable in total as outlined in Chapter 10, it can actually be carried out quite quickly and easily for a short course if the instructor is skillful and well prepared in the teaching of the course content.

It must also be remembered that any test, no matter how well designed, cannot measure all of the important learning outcomes for persons enrolled in a course. People have their own reasons for enrolling in courses and often have valuable learning outcomes unrelated to the formal objectives of the course. These types of individual, and sometimes unexpected, outcomes have been described in earlier chapters.

It is much more appropriate to evaluate the general effectiveness of courses in achieving specified learning outcomes across persons than it is to evaluate the learning of specific persons by tests or other assessment procedures. If the courses and programs offered by a college or university can generally be shown to meet their intended objectives, and if the people who operate these programs and courses can be determined to be responsible and competent on the basis of past performance, the claims made for future courses in advertisements are credible for future clients. Persons can enroll in courses of their choice for their own purposes. What they take away from the course at its conclusion, in terms of their own feelings of personal relevance, utility of course content, interests, new perceptions, and attitudes will almost always have much more meaning to them than any test score or set of test scores. The best use of tests is to evaluate the effectiveness of courses toward specific intended learning outcomes, not for the definitive determination of how much any one individual has learned from the experience of the course.

Making Course Evaluations Public

As has been mentioned in many places earlier in this book, the evaluation of courses must necessarily include much other information than the measured achievement of students based upon testing. The perceptions of participants enrolled in courses and their employers about the relevance, conditions, and quality of instruction; the competence of the instructor in the content of the course and in teaching; and records of the operating characteristics of courses must all be used in the evaluation of the effectiveness of courses and programs. Other information concerning the organization of the sponsoring program and the competence of its administration are also involved. The specific learning outcomes resulting from testing in specific aspects of course performance have little meaning without this additional contextual information. By itself, the performance data on specific appects of the course has little influence or utility. It is not by itself very convincing to persons who make decisions about enrolling or not enrolling in future courses. The tacit evaluation which governs these types of decisions is almost always based upon other types of information other than test scores.

All of these additional types of information ought to be collected routinely. This information, along with the achievement outcomes, similar to those presented in Figures 2 through 4, ought to be tabulated and presented for public examination. Companies and the individuals who are sent by companies to engage in learning through continuing education courses have a right to know the credentials of courses, programs, instructors, and institutions which offer continuing education programs. Individual engineers, who may wish to enroll in courses, and their professional societies also have a similar right to this type of information. Consequently, there is the need to collect and present in cogent tabular, graphic, and narrative form much additional information other than simply specific course learning outcomes if judgments of effectiveness of programs is to be made in a reasonable manner.

The routine collection of this comprehensive information about the operation of courses within continuing education programs can be very helpful, not only to clients and consumers of the courses, but to the institution which offers the courses. Demands for accountability are increasing. One of the best ways of being accountable is to systematically collect this range of information and make it widely and publically available to any persons wishing to examine it. The best interests of the public which consumes the courses and the institution offering courses and programs may be served in this manner.

Conclusion

This chapter has focused on the needs of various persons involved in continuing education courses and experiences in engineering to have good information concerning the degree to which intended learning outcomes have been achieved and the effectiveness with which courses have operated. Examples of how to persent and interpret achievement data based upon pre- and post testing in key performance areas of courses have been provided. Means for reporting the results of learning outcomes to individual students have been described as well as precautions for protecting the privacy of individuals and preventing the misuse of this information have been discussed. It is argued that carrying out these types of assessment activities is most useful for making decisions about the effectiveness of programs and courses, and least effective for making sharp distinctions between persons and how much each individual has learned following a course. It is also argued that responsible continuing education programs should consistently seek information about the achievement of their course enrollees, as well as much other information about the effectiveness of instruction and course operating characteristics. This information should be used for improving continuing education courses and programs. It should be summarized in cogent ways and be publicly. disseminated as part of the accountability process.

Chapter 14

RECOMMENDATIONS FOR EVALUATED CEUs

There is currently much interest in continuing education in engineering and other technical fields about "evaluated" continuing education units. The idea is that persons who complete continuing education courses need to be held accountable for having actually learned something. Certificates of attendance are not acceptable to many persons and groups as evidence of learning resulting from continuing education activities. It is this concern which has motivated many of the activities of The Learning Outcomes Measurement Project with its emphasis upon ways of measuring the learning resulting from short courses typical of those offered through continuing education programs.

Courses Where Tests Provide Accurate Estimates of Learning

The activities of the project staff have shown that there are a number of ways by which to measure and estimate the degree of learning resulting from a continuing education course. Some of the methods by which the learning outcomes of a course may be measured depend upon the type of course which is being considered. As has been noted earlier, the best measure of the learning resulting from a course designed to remediate or upgrade general knowledge and skill, such as is required to pass State licensing examinations, is the student's actual performance on the licensing examination. It is possible to sample a few items from the total domain of items on the professional examination which accurately estimate the individual's performance on the longer test. This can be accomplished through the use of relatively new psychometric test construction procedures involving the use of latent trait item analysis of test scores (Lord & Novick, 1968; Shoemaker, 1973).

Thus, for courses of this type it is possible to develop and use short but powerful tests to determine not only the learning outcomes by individual students after the course, but by which to determine which students need to take the course in the first place. Once such a test is developed for a remediation or upgrading of general knowledge and skills course, it can be used as a pre-test for advisement and screening purposes or as a post test for assessment of the learning of individuals who have completed the course. The scores of individuals on pre- and post tests over replications of the course may also be recorded, summed and the mean values and standard deviations calculated.

From this information it is also possible to evaluate the course and its effectiveness as well as to evaluate the learning of individual students. For such courses carefully constructed pre- and post tests are very useful and very accurate indicators of the degree of learning achieved by individual students and the general effectiveness of the course. The primary reason for this is that the performance domain of interest is clearly defined as a body of knowledge and skill thought to be basic to an area of engineering practice and incorporated on a broad spectrum professional examination.

Courses Where Tests are Inadequate Estimates of Learning

Evaluation of the learning outcomes of other types of continuing education courses is more difficult. This is because the domain of performance in which course knowledge and skill may be applied is not well defined. It is also because the ways in which the specific concepts and skills acquired in the course may be applied on the job by the engineer after the course is completed cannot be clearly and unambiguously stated. An illustration may help clarify this point.

Suppose a course is developed to broaden and update the skills of mining engineers in terms of the design of drainage and storage structures for the runoff for surface mined lands. The course presents the latest thinking, alterations in older theoretical models, and newly developed algorithms and nomographs by which to make accurate and efficient calculations of the design of these structures to meet newly developed Federal standards for water quality control downstream from the mining area. This is the type of course developed by Professors Haan and Barfield (1978) titled "Hydrology and

Sedimentology of Surface Mined Lands." This particular course was studied by the project team and tests were developed by which to measure entry level and exit level knowledge and skill of course participants in course content. The problem is that no matter how good the tests are they can never be a truly effective means of estimating all of the important learning outcomes which may have been achieved by any individual following completion of the course. There are several reasons for this.

Limited Time for Testing

First, in any continuing education course dealing with a large and complex body of information and skills, most of the time needs to be devoted to instructional activities. Students need to have concepts and procedures demonstrated and they need to apply these in practice problems. Most continuing education courses are short because practicing engineers cannot afford to spend long periods of time in course attendance. Thus, a short course of 2 to 4 days duration is a common occurrence. However, a truly adequate test of the learning outcomes of the Hydrology and Sedimentology course would require that the student actually construct the design specifications for a water drainage and storage system for an actual mining operation. This would usually require a minimum of from 6 to 8 hours depending upon the problem characteristics. There is simply not enough available time in the short course to devote such large amounts of time to a "test" item. This becomes particularly apparent when one realizes that a good test would require the actual design of, not one, but several drainage and storage systems for different topography, climatic, soil, and mining conditions.

Because of this problem of time, any test must be greatly abbreviated and simplified. This simplification and abbreviation makes the test tasks different from the performance tasks actually involved in the design of such structures on the job. This means that any good evaluation should require the engineer who has completed the course to submit a sample of his next few actual drainage and storage designs to the course instructor. These would be designs the engineer had actually produced on the job after the course had been completed. The course instructor could then evaluate the degree to which the principles and techniques taught in the course had been accurately applied and used by the engineer. Of course the problem is that it would take many hours for the instructor to evaluate each actual design, the evaluation could not be completed until some weeks after the student had completed the short course and had time to learn how to apply the course principles in the work setting, and it would take much more time and money than would possibly be available to complete such a thorough evaluation of each learner's achievement.

Inadequacy of Testing in Sampling the Performance Domain

A second problem is that even if such an elaborate evaluation of each person's learning were adopted, it would not be valid for some of the course participants. Rather, it would be valid for only those persons who came to the course with the intention of actually using the course principles and techniques in their daily work activities in specific design problems.

Experience of the project staff and many others has shown that for any particular course there are a wide variety of learners enrolled for a variety of reasons. For example, in the Hydrology and Sedimentology course, one often finds persons enrolled who are not normally engaged in the design of drainage and storage structures for surface mining. Sometimes persons attend such a course because they have business dealings with engineering firms which do carry out such designs. The purpose of attending the course is to become more informed about the problems and methods used by these designers and not to become expert in the actual design of the structures themselves. Administrators of state regulatory agencies, state inspectors, and other persons also not normally engaged in the actual design of such structures frequently attend such a course.

All of these persons may learn a great deal from the course but none of them might be expected to put into practice the actual principles taught in the course. The valuable things learned by this group of persons may be the names of research persons who they can hire as consultants to provide the technical assistance they need for specific jobs; the names of resource manuals and documents as well as computer programs useful to the solving of particular problems; and identification of areas of expertise and knowledge presently lacking in themselves or in members of their organizations which need to be developed either through the upgrading of present employees skills or the hiring of new employees.

All of these outcomes can be very valuable to the participants. They are certainly all outcomes which would be valued in terms of making a contribution to engineering practice in the region. Yet, none of these outcomes would be measured by a comprehensive evaluation of the actual designs for drainage and storage structures produced by this group of persons. Generally persons in this group would not produce such designs when they returned to work. Rather, they would use the knowledge they acquired to better manage their firm, supervise employees, and obtain the services and resources needed by the actual designers of such structures within their firms or under their jurisdiction.

Growth of Learning After Course Completion

A third problem has to do with when the learning resulting from the course may be expected to take place. In highly complex and technical courses which have large amounts of content, all that can be hoped for in a short course as an immediate outcome is a general familarity with the course concepts and procedures.

Actual facility in the use of these concepts and procedures is almost certainly dependent upon serious continued study and attempts by the course participant to actually apply and use the course material in his or her work setting. Therefore, the maximum amount of learning should not occur at the end of the short course but sometime after the completion of the short course when the learner has had time to do much additional further study and application of course procedures. Courses in computer programming are good examples. They usually teach only the basic principles and how to understand complex procedures and manuals. Facile computer programming comes only after much actual application of these concepts and principles in many work related problems.

The Need for Multiple Indicators of Learning Outcomes

For all of these reasons it is important to have multiple indicators of the degree of learning resulting from most continuing education courses if one wishes to determine the actual learning outcomes achieved. Simple pre- and post test scores of individuals at entry and exit from the course are useful but not sufficient to the task. Short entry and exit tests can provide good estimates of the degree to which participants have learned the basics of the course principles and procedures. If properly constructed they can also estimate the degree to which participants know how to go about using the materials, manuals, algorithms, nomographs, and other procedures presented in the course for approaching or setting up a few sample problems likely to be encountered in their actual work setting. Performance on such tests tells very little about other important outcomes the engineer may have learned.

This means that the learning outcome evaluation of a course's effect needs to include not only pre- and post tests on the basics of the course content, but also systematic polling of participants concerning what they think they have learned, how much they think they have learned, how relevant they think the learning is to their job performance, and how likely they are to do additional study in this area and attempt to apply course concepts and procedures to their actual work. The intentions and perceptions of course participants are very important!

The same is true for the perceptions of the supervisors and employers of the course participants. The judgments of these persons about the worth and utility of the material learned in the course by the employee is very important. It is the tacit evaluation of these supervisors as well as of the engineers who enroll in such courses which is the most potent and meaningful evaluation for any course.

If the judgment of these groups is that the course is valuable and worthwhile in terms of improving knowledge and skill in areas related to on-the-job performance, the course will be highly recommended and heavily subscribed whether or not there is any formal assessment of learning outcomes by testing or by the examination of actual job performance or work samples of course participants after the course is completed. If the tacit evaluation of these professional groups is that the course is not worthwhile, no matter how valuable the course is shown to be in improving test scores, it is not likely to be heavily enrolled. This professional judgment tacit evaluation is an important and legitimate part of the information which should be routinely gathered and incorporated in evaluations of the learning outcomes of continuing education courses.

The Impossibility of Making "Complete" Learning Assessments of Individuals

It should be apparent that it is an impossible task to collect all of this wide range of information in order to evaluate the learning of any one particular person who has taken a particular continuing education course. There is not time to do so and the entire process makes unreasonable demands upon the participant and his or her employer. How, then, can an individual enrollee in a given course be certified as having learned a specific amount in a given course?

Beyond certifying that the individual has attended, participated fully, completed all course assignments and activities at described levels of accuracy, and has also completed a pre- and post test which demonstrated a certain amount of growth on some of the basic knowledge and skill areas in the course, there is little that can be said about an individual's actual learning outcomes of a more broad and important nature which may result from the course. However, it is possible to evaluate the effectiveness of the course for participants generally in more substantial ways. In a nutshell, it is much more desirable to certify courses than persons.

Means for Making Comprehensive Assessment of Course Effectiveness

Although it is impractical to obtain all the various types of evidence needed to make a strong inference about the broad range of learning outcomes which may result from a particular person completing a course, it is practical to gather this wide array of evidence across different persons who have completed a course. While the instructor cannot hope to collect and evaluate the accuracy of the application of course principles to the actual design structures of practicing engineers for real problems from the field for each course participant, the actual designs from two or three persons enrolled in a course can be randomly solicited and evaluated. Over several replications of the course this practice reveals much information about the effects of the course in actually assisting the engineers enrolled in producing better structures. It also reveals much about the variation in the degree to which the course principles are appropriately applied following the completion of the course.

Other participants and their employers and supervisors can be sampled and interviewed about the actual degree they judge course principles and procedures are being used. Again, not every person must be interviewed.

Other persons from among the population of past course enrollees can be sampled and asked to complete a delayed post test of course content and skill. It is even possible and sometimes desirable to administer different test items or tasks to different persons at the end of a course. This would be done when there is a large amount of material to test, time only to administer a few test items to any one person, and an interest in learning something about the effectiveness of the course over the entire large array of items.

Such a plan used over several replications of a course produce much information about the course effectiveness in teaching its participants a variety of outcomes. Of course, it provides little information about any individual's learning.

Logical Requirements for Certification of Courses

Any movement toward an evaluated continuing education unit probably ought to be based on the course developers having to provide information about the general effectiveness of the course. This information should be based on the evaluation of actual on-the-job performance of samples of persons who have completed the course. It should also include the perceptions of samples of past enrollees and their supervisors concerning the value and utility of the course for improved performance in work related activities. This information should be presented along with information about entry and exit level knowledge and skill on the basics of the course as these can be measured in short and time efficient pre- and post tests. Professional licensing agencies and other interested groups such as practicing engineers and the firms which employ them ought to have access to this information. The effectiveness of the course could then be judged in a more formal way than the present and common tacit evaluation way, but without removing this valuable professional judgment component. Courses could be certified as being worthy of an evaluated CEU on the basis of this evidence.

Logical Requirements for Certification of Persons

The entry level knowledge and skills of participants, the reasons persons enroll in a particular course, their expectations for learning from the course, a record of their actual participation and completion of course activities, their actual performance on short and basic tests of key course knowledge and skills, and their perceptions of how much they have learned following the course, should all be rountinely collected for persons enrolled in continuing education courses. However, this assessment should be carried out only with the consent of the enrollee. Otherwise the results are likely to be invalid. If a course enrollee wants to receive an evaluated CEU, or some other type of certificate which reports learning, he or she should first be enrolled in an approved course which has demonstrated to the satisfaction of a professional licensing agency that the course does indeed achieve its intended learning outcomes in a consistent manner.

The second condition for earning an evaluated CEU or other formal credit should be that the individual participant be willing to complete short pre- and post tests and questionnaires designed to obtain basic information about individual learning which can be reliably and easily collected on each person in short periods of time at the beginning or end of the course. A third condition is that the participant engage in and complete all prescribed learning activities which comprise the course.

The Importance of Options for Participants

It is important that enrollees in a course be allowed the option of whether or not to receive CEUs or some other form of formal credit for learning. Some persons may be expected to be enrolled and not be at all interested in receiving formal documentation of their learning. However, if persons are interested in receiving such credit, they should be expected routinely to complete all pre- and post tests and related short questionnaires which elicit information about expectations and reasons for attendance, estimates of individual achievement, as well as judgments by participants about the utility of course content.

Persons wishing to receive formal credit also should be informed that they and some of their employers will be sampled in the future for follow-up interviews, delayed post testing, evaluation of job performance, and submission of actual work samples to be evaluated. It should be clear that the purpose of this follow-up assessment is for purposes of determining the general effectiveness of the course in reaching its intended learning outcomes in order that the course may be improved and eventually documented as being worthwhile for CEU credit.

The purpose of the follow-up assessment is not generally for making a judgment about the individual's learning for which personal CEUs would be awarded. Rather, successful completion of all course activities including the testing and learning assessments would usually be the basis for awarding individual CEU credit.

Involving All Participants in Learning Assessment Activities

This emphasis upon the participation of persons seeking CEU credit in the total assessment procedures for a course does not imply that other persons taking the course should not be involved. There is a need to assess the learning of course participants generally. Data gathered from across all persons enrolled is needed to improve the teaching of the course (formative evaluation) and to document the present effectiveness of a course in achieving its intended learning outcomes (summative evaluation).

Many engineers enrolled in short courses do not care about being awarded CEUs or other formal credit for their learning. If the testing procedures are presented *only* as being related to CEUs these individuals might opt to not participate in the assessment procedures. However, well designed assessment procedures also serve important instructional functions which are of benefit to all the persons enrolled in a course. This relationship of testing to instruction is particularly obvious in the embedded test tasks in the course which are used to inform the learner and the course instructor of the needs and accomplishments of the individual during the course of instruction in order to make instructional decisions (See Chapter 7). Quizzes, homework problems, and laboratory exercises are typically used for this purpose.

To make these activities optional would be to remove an important instructional component of the course for some persons. Generally these activities should be required for all participants because they comprise an integral part of the course and its instructional methods.

The same relationship should hold for other types of testing and assessment procedures as well. Whatever other purposes they serve, tests and other assessment procedures should _always_ serve instructional purposes. If tests and assessment procedures are developed in the manner suggested in Chapter 10 and other sections of this book, this will be the case. In this event participation in the pre- and post testing as well as all other assessment procedures ought to be built in as part of the regular instructional activities in the course. All participants should be involved in these activities. The results will not only be useful to the improvement of the course and the documentation of its general effectiveness, but will aid the learning of individual participants in a number of ways. Pre-tests inform the learner about the specific content and objectives of the course in a very precise way. This information is useful to the individual in focusing attention on relevant aspects of the course material during instruction. Post tests, when compared to pre-test results, inform the individual learner about his or her progress through the course in specific topics and areas and call attention to areas in need of additional study. Individuals are almost always interested in the growth of their own level of knowledge and skill and in comparing their progress with the accomplishments typical of other participants in the course as well as to persons in other sections of the same course taught at other times. Even when persons have no desire to receive CEU credit they remain interested in information about their own degree of learning and accomplishment.

Maximum participation of individuals enrolled in a course in the learning assessment procedures can be insured if the learning assessment procedures are fully integrated with the instructional procedures and if a number of policies are followed. The testing and other assessment procedures should be abbreviated and time efficient. All tests should be valid and reliable. Results of tests and other learning assessments should be shared as soon as possible with the participants. Individual test results should be kept private and not shared with others without the specific permission of the individual.

The uses of the learning assessment data gathered for the purposes of formative evaluation and documentation of course effectiveness ought to be explained to participants in order that they fully understand the importance of their participation in and contribution to the evaluation of the course. When these procedures are followed all participants will quite naturally be involved in the assessment procedures. They, the course instructors, and groups of persons who will become enrolled in the course in the future will all benefit.

Conclusion

There is no simple way to evaluate complex learning outcomes which may be expected to result from the completion of most continuing education courses in technical fields. If professional agencies and organizations are serious about developing evaluated CEUs, procedures similar to those described in this book and summarized in this chapter will need to be developed and followed.

REFERENCES

Airasian, P. W. & Madaus, G. F. Criterion-referenced testing in the classroom. In R. W. Tyler & R. M. Wolf (Eds.), Crucial issues in testing. Berkeley, California: McCutchen, 1974.

Aleamoni, L. M. Evaluation as an integral part of instructional and faculty development. In L. P. Grayson & J. M. Biedenbach (Eds.), *Proceedings 1980 industry education conference*. Washington, D.C.: American Society for Engineering Education, 1980, 120-123.

American Educational Research Association monograph series on curriculum evaluation. Volumes 1 through 5. Chicago: Rand McNally, 1967, 1968, 1969, 1970, 1970.

Bellack, A. A. & Kliebard, H. M. (Eds.). *Curriculum and evaluation*. Berkeley, California: McCutchen, 1977.

Bloom, B. J. *Human characteristics and school learning*. New York: McGraw Hill, 1976.

Bloom, B. S. (Ed.). *Taxonomy of educational objectives: Handbook I: Cognitive domain*. New York: David McKay Company, Inc., 1956.

Bloom, B. S., Hastings, J. T., & Madaus, G. F. *Handbook on formative and summative evaluation of student learning*. New York: McGraw Hill, 1971.

Box, G. E. P., Hunter, W. C. & Hunter, J. S. *Statistics for experimenters: An introduction to design, data analysis, and model building*. New York: John Wiley & Sons, 1978.

Bugelski, B. R. *The psychology of learning applied to teaching* (2nd ed.). Indianapolis: Bobbs-Merrill, 1971.

Carroll, J. B. A model of school learning. *Teachers College Record*, 1963, *64*, 723-733.

Cleaver, T. G. A controlled study of the semi-paced teaching method. *Engineering Education*, 1976, *66*, 323-325.

Cole, H. P. *Process education*. Englewood Cliffs, New Jersey: Educational Technology Publications, 1972.

REFERENCES

Cole, H. P. Principles and techniques for enhancing motivation and achievement of engineering students. In L. P. Grayson & J. M. Biedenbach (Eds.), Proceedings 1980 College Industry Education Conference. Washington, D.C.: American Society for Engineering Education, 1980, 345-347.

Council on the Continuing Education Unit. The continuing education unit: Criteria and guidelines. Silver Springs, Maryland: The Council on the Continuing Education Unit, 1979.

Enell, J. W. The CEU in the 1980's: A report from a long term user. In L. P. Grayson & J. M. Biedenbach (Eds.), Proceedings 1980 Industry Education Conference. Washington, D.C.: American Society for Engineering Education, 1980, 185-189.

Ericksen, S. C. Motivation for learning: A guide to the teacher of the young adult. Ann Arbor, Michigan: University of Michigan Press, 1974.

Ferry, R. Tests in adult non-credit education. CPD2 Newsletter, Fall, 1979.

Gage, N. L. & Berliner, D. C. Educational psychology. Chicago: Rand McNally, 1975.

Gagne, R. M. The acquisition of knowledge. Psychological Review, 1962, 69, 355-365.

Gagne, R. M. The psychological basis of science -- A process approach. Washington, D.C.: American Association for the Advancement of Science, Commission on Science Education, 1965.

Gagne, R. M. Curriculum research and the promotion of learning. In R. Tyler, R. Gagne, & M. Scriven (Eds.), Perspectives of Curriculum Evaluation. Chicago: Rand McNally, 1967, 19-38.

Gagne, R. M. The conditions of learning (3rd ed.). New York: Holt, 1977.

Gagne, R. M. & Briggs, L. J. Principles of instructional design. New York: Holt, 1974.

Gagne, R. M. & Paradise, N. E. Abilities and learning sets in knowledge acquisition. Psychological Monographs, 1961, 75 (whole No. 518).

REFERENCES

Greenfield, L. B. Comments on evaluation in engineering education. *Engineering Education*, 1978, *68*, 401-404.

Grobman, H. *Evaluation activities of curriculum projects*. Chicago: Rand McNally, 1968.

Grogan, W. R. Performance-based engineering education and what it reveals. *Engineering Education*, 1979, *69*, 402-405.

Haan, C. T. & Barfield, B. J. *Hydrology and sedimentology of surface mined lands*. Lexington, Kentucky: Office of Continuing Education and Extension, College of Engineering, University of Kentucky, 1978.

Hambleton, R. K. & Novick, M. R. Toward an integration of theory and method for criterion-referenced tests. *Journal of Educational Measurement*, 1973, *10*, 159-170.

Heimback, C. L. To PSI and back. *Engineering Education*, 1979, *69*, 399-401.

Holland, J. L. *Making vocational choices: A theory of careers*. Englewood Cliffs, New Jersey: Prentice Hall, 1973.

Hoyt, D. P. The relationship between college grades and adult achievement. A review of the literature, *Research Report No. 7*. Iowa City, Iowa: American College Testing Program, 1965.

Klus, J. P. & Jones, J. A. *Engineers involved in continuing education: A survey analysis*. Washington, D.C.: American Society for Engineering Education, 1975.

Knowles, M. S. *The modern practice of adult education: Andragogy versus pedagogy*. New York: Association Press, 1970.

Kulik, J. A. & Kulik, C. C. Effectiveness of the personalized system of instruction. *Engineering Education*, 1975, *65*, 228-231.

Kulik, J. A., Kulik, C. C. & Cohen, P. A. A meta-analysis of outcome studies of Keller's personalized system of instruction. *American Psychologist*, 1979, *34*, 307-318.

Lacefield, W. E. *The evaluation of competence: Theoretical and empirical perspectives*. Unpublished doctoral dissertation, University of Kentucky, 1980.

Lavin, D. E. *The prediction of academic performance: A theoretical analysis and review of research*. New York: Russell Sage Foundation, 1965.

REFERENCES

Livingston, S. A. Criterion-referenced applications of classical test theory. Journal of Educational Measurement, 1972, 9, 13-26.

Livingston, S. A. A note on the interpretation of the criterion-referenced reliability coefficient. Journal of Educational Measurement, 1973, 10, 311.

Lord, F. M. & Novick, M. R. Statistical theories of mental test scores. Reading, Massachusetts: Addison-Wesley, 1968.

Manning, W. H. The criterion problem. In P. H. DuBois & G. D. Mayo (Eds.), Research strategies for evaluation training. Chicago, 1970, 68-78.

Maratuza, V. R. Applying norm-referenced and criterion referenced measurement in education. Boston: Allyn and Bacon, 1977.

Marion, R. An evaluation model for developmental growth. Unpublished doctoral dissertation, University of Kentucky, 1978.

Marshall, J. C. & Hales, L. W. Classroom test construction. Reading, Massachusetts: Addison-Wesley Publishing Co., 1971.

Marshall, J. C. & Hales, L. W. Essentials of testing. Reading, Massachusetts: Addison-Wesley, 1972.

Martin, E. D. & Greenfest, M. C. Criteria and standards: An institutional evaluation model for CEU activities. Paper presented at Lifelong Learner Research Conference, University of Maryland, February 1, 1980. Richmond, Virginia: Virginia Commonwealth University.

Mason, E. J. & Bramble, W. J. Understanding and conducting research/Applications in education and the behavioral sciences. New York: McGraw-Hill, 1978.

McClelland, D. C. Testing for competence rather than for intelligence. American Psychologist, 1973, 20, 1-14.

McCullough, R. C. Current research on roles and competencies of professional trainers. In L. P. Grayson & J. N. Biedenbach (Eds.), Proceedings 1980 industry education conference. Washington, D.C.: American Soceity for Engineering Education, 1980, 282-292.

REFERENCES

Mertens, D. M. *Results of AESP's Survey of Program Interests for Engineers*. Unpublished paper. Appalachian Education Satellite Program, University of Kentucky, 1978.

Miller, D. B. *Personal vitality*. Reading, Massachusetts: Addison-Wesley Publishing Co., 1977.

Millman, J. Criterion-referenced measurement. In W. J. Popham (Ed.), *Evaluation in education, current applications*. Berkeley, California: McCutchen, 1974.

Morris, A. J., Sherrill, P., & Scriven, M. *The return on investment in continuing education of engineers*. (Research report based on work partially supported by National Science Foundation, Grant No. EPP75-21587, June 1978).

Morstain, B. R. & Smart, J. C. Reasons for participating in adult education courses: A multivariate analysis of group differences. *Adult Education*, 1974, *24*(2), 83-98.

Moss, P. J. Barfield, B. J., & Blythe, D. K. Evaluation in continuing education: A pilot study. In L. P. Grayson & J. M. Biedenbach (Eds.), *Proceedings 7th annual frontiers in education conference*. Washington, D.C.: American Society for Engineering Education and Institute of Electrical and Electronics Engineers, 1977, 337-344.

Nader releases ETS report, hits tests as poor predictors of performance. American Psychological Association, *APA Monitor*, 1980, *11*, 30-31

Nunnally, J. C. *Educational measurement and evaluation*. New York: McGraw Hill, 1972.

Salvendy, G. & Seymour, W. D. *Prediction and development of industrial work performance*. New York: Wiley, 1973.

Scriven, M. The methodology of evaluation. In R. Tyler, R. Gagne, & M. Scriven (Eds.), *Perspectives of curriculum evaluation*. Chicago: Rand McNally, 1967, 39-83.

Shoemaker, D. M. *Principles and procedures of multiple matrix sampling*. New York: Ballinger, 1973.

Snelbecker, G. E. *Learning theory, instructional theory, and psychoeducational design*. New York: McGraw Hill, 1974.

REFERENCES

Stice, J. E. Grades and test scores: Do they predict adult achievement? *Engineering Education*, 1979, *69*, 390-393.

Thorndike, R. L. (Ed.). *Educational measurement* (2nd ed.). Washington, D.C.: American Council on Education, 1971.

Tyler, R. W. *Constructing achievement tests*. Columbus, Ohio: Ohio State University, 1934.

Tyler, R. W. *Basic principles of curriculum and instruction*. Chicago: University of Chicago Press, 1950.

Tyler, R. W. The use of tests in measuring the effectiveness of educational programs, methods, and instructional materials. In R. W. Tyler & R. M. Wolf (Eds.), *Crucial issues in testing*. Berkeley, California: McCutchen, 1974.

Tyler, R. W. & Wolf, R. M. *Crucial issues in testing*. Berkeley, California: McCutchen, 1974.

Webb, W. B. Measurement of learning in extensive training programs. In P. H. DuBois & G. D. Mayo, *Research strategies for evaluating training*. Chicago: Rand McNally, 1979, 55-65.

Wiesehuegel, R. E. *Measurement of cognitive achievement in continuing education in engineering*. Unpublished doctoral dissertation, 1978, George Peabody College for Teachers.

Wolf, R. M. Invasion of privacy. In R. W. Tyler & R. M. Wolf (Eds.), *Crucial issues in testing*. Berkeley, California: McCutchen, 1974.

Work, C. E. A nationwide study of the variability of test scoring by different instructors. *Engineering Education*, 1976, *66*, 241-248.

Worthen, B. R. & Sanders, J. R. *Educational evaluation: Theory and practice*. Belmont, California: Wadsworth, 1973.

APPENDIX A

This appendix contains an example of four different types of data collection instruments suitable for use in educational and evaluational situations involving short courses. A brief description of each instrument - its purpose, its recommended implementation mode, and its possible utilities - is given.

The four different types of instruments serve to provide additional information about the participants, instructors, and operating characteristics involved in courses. This type of descriptive information is needed to properly interpret the results of formal assessments of participants' learning by testing. The sample instruments presented here, used in conjunction with the sample learning assessment tests in Appendix B, collectively allow strong judgments to be made concerning the effectiveness of courses in achieving intended objectives (summative evaluation) and in reorganizing courses to be more effective in the future (formative evaluation).

APPENDIX A

EXAMPLE A-1

Instrument:	Demographic Information Questionnaire
Purpose:	A) To collect systematic data concerning participants' personal, educational, and employment histories. B) To collect data concerning the relative influence of a number of factors affecting decisions to participate in a particular continuing education program.
Implementation:	Either A) As part of an advance mailing of materials to be completed and returned by participants through the mail or upon arrival at the course site. Or B) Completed by participants as an initial activity during the first formal meeting of the course.
Utilities:	A) To identify characteristics of the "captured" audience for contrast with those of the intended "target" audience for the course. B) To provide a source of information reflecting on the validity of evaluation methods and outcomes concerning course effectiveness. C) To aid faculty in selecting course content and designing appropriate instructional methods. D) To aid sponsors to identify topics and plan advertising methods for future courses.

COURSE: ____________________

INSTRUCTOR(S): ____________________

DEMOGRAPHIC INFORMATION

	Items	Responses	Comments
I.	In what type of engineering are you currently employed? 1. Agricultural 2. Chemical 3. Civil 4. Electrical 5. Industrial 6. Mechanical 7. Mining 8. Other(PS)* 9. Not presently employed as an engineer(PS)*	1 2 3 4 5 6 7 8 9	
II.	What is your highest educational degree? Please state major field in comments. 1. High School 2. Associate 3. Bachelor's 4. Master's 5. Doctorate 6. Other(PS)*	1 2 3 4 5 6	
III.	What is your sex? 1. Female 2. Male	1 2	
IV.	What is your major employment affiliation? 1. University or college (not a student) 2. Government 3. Consultant 4. Corporate (in business or industry) 5. Student 6. Unemployed 7. Other(PS)*	1 2 3 4 5 6 7	
V.	Who is paying for your attendance at this course? 1. My employer 2. My self 3. Other(PS)*	1 2 3	
VI.	Did your employer recommend this course? 1. No 2. Yes	1 2	
VII.	How did you hear about this course? 1. Brochure/posted 2. Brochure/mailed 3. Word of mouth 4. Newspaper 5. Professional journal 6. Radio or TV 7. Other(PS)*	1 2 3 4 5 6 7	
III.	What is your age? 1. Less than 21 years old 2. 21-30 years old 3. 31-40 years old 4. 41-50 years old 5. More than 50 years old	1 2 3 4 5	
IX.	What is your race? 1. Caucasian 2. Black 3. Oriental 4. American Indian 5. Other(PS)*	1 2 3 4 5	

*PS=Please specify in comments.

Demographic Information Questionnaire

Example A-1

Example A-1 continued

Demographic Information Questionnaire

X. Have you previously attended this course or other continuing education courses in this subject area?

1. No 2. Yes (Please list the three most recent) 1 2

Course Institution

1.
2.
3.

On a scale from 1 to 5 (1=very important; 5=unimportant), please rate how important the following factors were in your decision to attend this course. Circle your response. If the factor is not applicable to your situation, circle number 6.

		Very Important				Unimportant	Not Applicable
XI.	My employer recommended the course.	1	2	3	4	5	6
XII.	I was interested in the subject area.	1	2	3	4	5	6
XIII.	My expenses were paid.	1	2	3	4	5	6
XIV.	The host institution and/or instructor(s) were noted for their expertise in this subject area.	1	2	3	4	5	6
XV.	I have previously attended this and/or similar courses, and have found them to be of value.	1	2	3	4	5	6
XVI.	I wanted to meet and exchange ideas with my colleagues.	1	2	3	4	5	6
XVII.	I need this course to maintain my present position or to be considered for a promotion.	1	2	3	4	5	6
XVIII.	I wanted to learn or refresh my knowledge and skills in this subject area, so my job performance may be enhanced.	1	2	3	4	5	6

XIX. Please list any other factors which influenced your decision to attend this course.

APPENDIX A

EXAMPLE A-2

Instrument: Participant Reaction Questionnaire

Purpose:

A) To collect impressions of participants regarding:

1) The course faculty as instructors.
2) The course content and presentation mode(s).
3) Specific learning outcome characteristics.
4) Anticipated usefulness of knowledge and skills acquired through the course.

B) To elicit participants' general perceptions and comments concerning the course.

C) To identify further topics and areas of interest to participants.

Implementation: To be administered at or near the end of the final formal meeting of the course, either before or after any posttest. (After the posttest, after discussion of that test, and prior to or during closure is a recommended time for this administration.)

Utilities:

A) To evaluate effectiveness of the course in terms of several broad process indicators.

B) To provide a source of formative information feedback to instructors concerning participant perceptions of content, instructional process, and value of the course.

C) To aid sponsors to evaluate audience receptivity to the course, its content, and its faculty.

D) To provide information to future potential participants regarding the perceptions of previous participants.

Example A-2
Participant Reaction Questionnaire

COURSE: ______________________________

INSTRUCTOR(S): ______________________________

EVALUATION OF COURSE

On a scale from 1 to 5 (1=strongly agree; 5=strongly disagree), please respond to the following aspects of the course. Circle your response. Instructor(s)	Strongly Agree				Strongly Disagree	Instructor
1. The instructor(s) was knowledgeable in the subject area.	1	2	3	4	5	________
	1	2	3	4	5	________
2. The instructor(s) effectively communicated the knowledge and skills presented in the course.	1	2	3	4	5	________
	1	2	3	4	5	________
3. The instructor(s) was not receptive to your comments and needs.	1	2	3	4	5	________
	1	2	3	4	5	________
4. The instructor(s) use of examples and practice problems was effective in demonstrating the knowledge and skills presented in the course.	1	2	3	4	5	________
	1	2	3	4	5	________
Content & Presentation						
5. The text and/or reference materials were appropriate and useful.	1	2	3	4	5	
6. The organization of the course content was poor.	1	2	3	4	5	
7. The course content was not relevant to your work activities.	1	2	3	4	5	
8. Tutorial sessions, made available during the course, would be valuable additions to the structure of the course.	1	2	3	4	5	
On a scale from 1 to 5, please respond to items 9 and 10.	too difficult				too easy	
9. The level of difficulty of the material presented was:	1	2	3	4	5	
	too slow				too fast	
10. The rate of presentation of the material was:	1	2	3	4	5	
11. Comments on the course instructor(s), content, and/or presentation.						

Example A-2 continued

Participant Reaction Questionnaire

Course Objectives: To disseminate current information from various institutions, including universities, government agencies and practicing engineering groups, for the advancement of knowledge concerning these subjects and for implementation by field engineering applications.

	Learning Outcomes	Strongly Agree				Strongly Disagree
12.	The course met its stated objectives.	1	2	3	4	5
13.	The knowledge and skills I obtained from this course will be of no value to me in my job.	1	2	3	4	5
14.	The knowledge and skills I obtained from this course would have been difficult to obtain elsewhere.	1	2	3	4	5
15.	I met persons, other than the instructor(s) from whom I obtained valuable knowledge and/or information.	1	2	3	4	5
16.	I would not recommend this course to others in my position.	1	2	3	4	5
17.	It is likely that, in the future, I will contact one or more persons, whom I've met at this course, concerning some aspect of my work.	1	2	3	4	5
18.	I feel confident I can properly use the knowledge and skills I obtained through this course.	1	2	3	4	5
19.	I believe it is not appropriate to award CEU credit for this course.	1	2	3	4	5
20.	I feel confident in the validity of the results obtained from the techniques presented in this course.	1	2	3	4	5
21.	I intend to further my study into the subject area presented in this course.	1	2	3	4	5
22.	My experiences in the course were interesting and enjoyable.	1	2	3	4	5
	On a scale from 1 to 5 (1=very knowledgeable; 5=not knowledgeable), please answer questions 23 and 24.					
23.	How knowledgeable of the course content were you prior to entering the course?	1	2	3	4	5
24.	How knowledgeable of the course content are you now upon completion of the course?	1	2	3	4	5

25. Do you intend to share the knowledge and skills you obtained in the course with your colleagues at your place of employment?

    a) No
    b) Yes, I am required to do so by my employer. (How?) ______
    c) Yes, I intend to, although I am not required to do so by my employer. (How?) ______

    ______

26. What aspects of the course do you feel were:

    Most beneficial -

    Least beneficial -

27. Comments/suggestions concerning any aspect of the course.

28. Please list three (3) topics which you would like presented in a course that would be of value to you in your work.

    1.
    2.
    3.

APPENDIX A

EXAMPLE A-3

Instrument: Satisfaction/Utilization Survey Form

Purpose: To gather impressionistic and factual data regarding the applicability of the course content and materials vis-a-vis the participants' job roles and responsibilities.

Implementation: As a part of a follow-up study of short course participants. To be included among follow-up materials sent to participants three to six months after course completion.

Utilities:

A) As part of an extended course evaluation procedure and used in conjunction with data obtained prior to and during the course, this instrument provides data regarding the extent to which overall course objectives have been obtained.

B) Allows for correlational studies and validation of other evaluation information sources: e.g., in contrast to demographic data, facilitates the identification of characteristics of the target audience that will benefit most from the course.

C) Provides sponsors and others with documenting evidence of course value and applicability to participants in their work.

Example A-3

Satisfaction/Utilization Survey Form

---

1) Was the course a worthwhile professional experience for you?

DEFINITELY YES) _____
YES, MODERATELY SO) _____
NO) _____ A comment?) _____________________
____

2A) Would you recommend this particular course to other professionals who work in areas similar to your own?

YES) _____
NO) _____

2B) If YES to (2A), have you in fact done so?

YES) _____ How many times?) _____
NO) _____

3A) Please indicate the extent to which you have found the course subject matter and materials applicable to your work.

_____ I have found little or no relation between the course content and my normal work.

_____ I have referred to the materials and information presented during the course on several occasions since.

_____ The course content has proven moderately useful to me in my work. I refer to that content almost monthly.

_____ I have found the course content has extensive application in my work area and I refer to that content frequently, perhaps on a weekly basis.

3B) Rather than judging the value of the course in terms of the extent of its applicability and usefulness, have there been one or more occasions since you attended the course where the content and/or materials have been critical or otherwise very valuable to you in some phase of work on a particular problem, experiment, project, plan, or other activity?

NO) _____ YES) _____ How many such occasions? _____

If YES, would you comment briefly on the nature of these applications?

____

____

(Thank you)

APPENDIX A

Example A-4

Instrument: Structured Personal Interview Protocol

Purpose: To gather factual and impressionistic data concerning aspects of short course implementation in an industrial or business environment.

Implementation: The protocol provides a structural and substantive format to guide discussion during a personal interview between course designers or evaluators and knowledgeable representatives of corporate clients: e.g., a continuing education coordinator, a plant training manager or supervisor, or a course instructor.

Utilities:

A) To discover corporate perceptions of course utility and to identify further educational needs.

B) To clarify the characteristics of corporate personnel who comprise the course audience.

C) To identify preferred instructional procedures and presentation modes followed by corporate clients to implement the course (and to contrast these with design specifications for the course).

D) To identify formal and informal modes of course evaluation and assessment of participant learning presently employed in corporate settings.

E) To gather information required for formative and summative course evaluation.

F) To make acquaintances and friends in the field of potential course users and to test the marketability of new courses or course modifications.

Example A-4

Structured Personnel Interview Protocol

Questions about matters of fact and procedure:

I) Why did the company want or need a course like "Design of Experiments"?

II) How did the company find out about and select "Design of Experiments"?

III) What company personnel took the course?

- A) Why were these persons selected?
- B) How were these persons selected?
- C) What incentives, benefits, or compulsions were used?
- D) What consequences for subsequent employment and/or advancement opportunities were contingent upon successful course completion?
- E) What arrangements were made for:
  - 1) course time (i.e., regular duty, release time, off-duty, etc.)
  - 2) travel, meals, expenses, etc.
  - 3) books, materials, supplies, etc.

IV) How was the course implemented by the company?

- A) Scheduling meetings, films, discussions, examinations, etc.?
- B) Homework and other course-related activities outside formal class meetings?
- C) Was an experienced statistical consultant on hand to help students?
  - 1) Was this person a company employee or an outsider?
  - 2) Was there a consultant plan to utilize this person?
    - a) Did he grade or otherwise comment on homework?
    - b) Did he give demonstrations; act as an instructor?
    - c) Did he provide students with company/job related examples?

Example A-4 continued

V) How did the company evaluate the course and the students?

   A) The course costs money; what benefits accrued the company? How were these measured or appraised?

   B) Did the students complete "participants questionnaires" or other similar instruments or surveys?

      1) Were student "comments" solicited by course faculty and/or training school staff?

      2) Were these and/or other data sources used to:

         a) justify course expense?

         b) modify course design and implementation strategy?

   C) Were any formal achievement tests given students?

      1) Was a certain score on such a test used to indicate successful or unsuccessful completion of the course?

      2) Did such "grades" go into students' personnel files as permanent records?

   D) Were students' supervisors provided formal or informal reports?

Questions about matters of opinion:

I) Did students like the course? Did they think it was a worthwhile expenditure of time, effort, and money?

II) Is the company satisfied with the course as a whole?

   A) What features were especially good from the company's viewpoint?

   B) What needed to be or was done differently?

III) Could or should students be selected differently?

IV) Was a statistical consultant important? To what degree was the course "self-instructional"?

V) Would the company be interested in a formal evaluation process aimed at measuring and reporting individual student achievement?

   A) "Design of Experiments" costs about $120.00/student. Would the company be interested enough in formal evaluation to pay an additional $10-15 per student for this service?

APPENDIX B

SAMPLE ABBREVIATED EMBEDDED TEST FOR COMPREHENSIVE ASSESSMENT OF COMPLEX KNOWLEDGE AND SKILLS

The short test which follows is an actual test for one unit in a six unit course titled, Hydrology and Sedimentology of Surface Mined Lands, by C. T. Haan and B. J. Barfield, University of Kentucky: Office of Continuing Education and Extension, College of Engineering, 1978.

The course is a very popular short course taught in three day intense workshops. The enrollees are mining engineers and others with interests in the construction of better water drainage and storage structures for surface mining operations. The course is highly technical and develops an ability for participants to use a complex set of procedures presented in the course manual in the design of actual structures under very different types of slope, soil, climatic, and mining conditions.

The solution of entire real problems takes several hours and sometimes even a day or two. Therefore, the actual teaching of the course, as well as the testing of competence of participants at the end of a unit of instruction or the end of the course, cannot be based upon having participants complete actual entire problems. There simply would not be enough time.

The sample short test presented is one way to assess the knowledge and skill of course enrollees in the complex content of the course. The test items range from simple and basic understanding of principles through the application of these to the solution of realistic, complex problems. Persons' test scores reveal much about what the individual has learned and what he or she may not have learned.

Similar short tests may be constructed for other units in this course or other courses. These unit tests can be assembled into one comprehensive test. For this course it would take about one hour to complete such a 50 item test. The test would abbreviate a set of realistic problems which, if presented in full, would take many hours to complete. The test is, thus, an efficient estimate of the learning of persons based upon a much shorter time period of activity, provided the items are sampled appropriately and properly constructed.

As such a test is developed, parallel forms can be produced. This allows the use of short but comprehensive tests for pre-tests, embedded tests, post tests, and delayed post tests. All of these types of tests can be useful in assessing, not only the learning outcomes for a course for individuals, but for judging the effectiveness of the course as well.

Although abbreviated tasks of the type included in the sample test are never a substitute for the assessment of learning by observing actual on-the-job performance after the completion of a course, or by analysis of actual work samples of persons completed after the course, the abbreviated test tasks can be an efficient way to judge the degree of learning resulting from a course at its conclusion.

The sample items which follow, the explanation, commentary, and the guidelines which are included may be helpful to understanding how such efficient but brief tests of complex performances may be developed and assembled. Studying Chapters 7 and 10 will also add to this understanding.

The charts, tables, and nomographs which are attached to the sample test are taken from the Haan and Barfield (1978) manual. They contain information needed to solve the problems presented in the items. Coupled with the test items they test for the ability of individuals to make proper use of the manual and its materials in the solution of problems of a realistic nature. These realistic problems are presented in the test items. They are sampled from the domain of real problems frequently encountered in the design of open channel hydrologic drainage structures in surface mining situations. For convenience the complete sample test is presented in their appendix, although it occurs earlier in Chapter 10 as Table 5.

Performance Objectives for Which Items Were Written

As pointed out in Chapter 10, the specific performance objectives stated in operational terms need to be developed prior to the preparation of the instructional activities or test items by which to assess the achievement of these expected outcomes. The performance objectives for the open channel hydraulic structures unit of the Hydrology and Sedimentology course are stated in Table 4 in Chapter 10. For convenience this table of objectives is also presented in this Appendix. It is these particular performance objectives that the sample test items are designed to assess. The reader should now examine the performance objectives in Table 4, the test items developed to assess these objectives in Table 5, and then read the additional comments which follow. These explain the details of how each item operates, what it is intended to measure, and why. The example should be useful to persons wishing to construct similar tests for units in technical courses.

Presenting the Stimulus Elements Required for Performance

Appended to the set of test items students receive is a set of figures, charts and tables (Figure 8). One main objective of the course is to teach students the proper use of these materials contained in the manual. All of the figures and tables appended to the test booklet have to be used to solve the problems or answer the questions, except for Figure 3.10. Since all the figures occur in one place with the tables after the test items, students have to discriminate from among the entire array the particular table or figure needed for a particular aspect of a problem.

Table 4

Performance Objectives for Open Channel Hydraulic Structures Unit: An Illustration of Test Construction Procedures*

Objective Number	Action Verb(s)	Description of the Performance Required and the Conditions Under Which it is to Occur
1	Describe	What happens to the value of Manning's n when the boundary of a channel varies through a range of structural conditions including different types of vegetation, non-vegetated soil aggregates, and man-made lining materials.
2	Recall, Recognize	The typical profile of flow velocities (fps) for hydrologic channels of various cross section shapes at typical slopes.
3	Describe Adjust Calculate	The relationship between retardance and flow rate in an hydrologic channel and make adjustments in design specifications (depth, top width, hydraulic radius, slope, and cross section) to produce desired freeboard and channel performance given changes in retardance or flow rates.
4	Calculate	By the limiting velocity method the permissible flow rate for channels given various slopes, required capacities, boundary conditions, soil types, and channel cross sections.
5	Calculate	By appropriate methods and proper use of tables and charts provided, the value of Manning's n for any type of channel given the boundary characteristics.

---

*See Appendix B for details about how the performance descriptions were developed and how test items were designed to measure each objective.

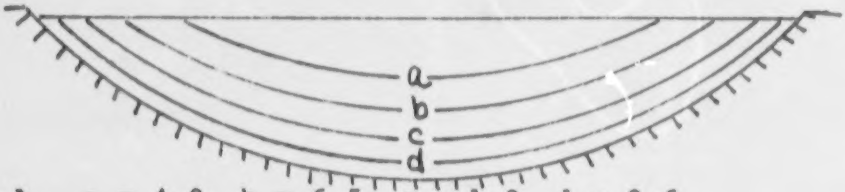
Table 4 (continued)

Objective Number	Action Verb(s)	Description of the Performance Required and the Conditions Under Which it is to Occur
6	Calculate	The hydraulic radius of channels of differing cross sections according to the appropriate modification of the basic computational algorithms.
7	Calculate	The design specifications for any given channel including the values Vp, R, S, D, T, and necessary freeboard given the specifications for any two of these values and information about soil type, topography, etc.
8	Design, Diagram, Label	A hydrologic channel designed to perform to stated specifications under stated problem conditions similar to those listed in item g above.
9	Recognize	The reasonableness of design specifications obtained as the solution to a particular design problem involving a hydrologic channel given the problem variables.
10	Use Select Doublecheck	Appropriately, computational short cut procedures, computational algorithms, and graphic solutions to complex equations given a variety of problems involving the design of hydrologic channels under widely differing conditions of rainfall, soil type, slope, etc.

Table 5

TEST FOR "OPEN CHANNEL HYDRAULICS" UNIT - Illustrating the Mapping of Items to Performance Objectives

1. What is a typical profile of flow velocities (fps) for the channel cross section represented in this figure?



   A. a = 4.9, b = 6.5, c = 1.2, d = 2.6
   B. a = 1.2, b = 2.6, c = 4.0, d = 6.5
   C. a = 6.5, b = 4.9, c = 2.6, d = 1.2
   C. a = 6.5, b = 6.2, c = 2.6, d = 2.3

2. What happens to the value of Manning's n when an erodible parabolic cross section open channel is vegetated compared to an identical nonvegetated channel?

   A. increases
   B. decreases
   C. remains unchanged
   D. varies with runoff volume

3. A nonvegetated trapezoidal channel through a sandy loam collidal soil has originally been designed to carry 8 cfs of water down a 4% slope. Suppose the engineer later decides to use a vegetated channel. What must he do to insure an equivalent capacity with the vegetated channel given the same slope, soil conditions, and channel shape?

   A. Select a grass which will grow to a uniform height without clumping to assure uniform flow rates at the channel perimeter.
   B. Design a somewhat deeper and wider channel to allow for the increased retardance of the flow caused by the vegetation.
   C. Design a somewhat shallower and narrower channel because with vegetation a higher flow rate can be sustained.
   D. Maintain the original specifications for the non-vegetated channel because the flow capacity will remain nearly unchanged.

Table 5 (continued)

* A channel is to be designed to carry 11.6 cfs of clear water down a 7% slope. The channel material is shale and hardpan. The channel is to be trapezoidal with a 3:1 side slope. Use this information to answer questions 4-8.

4. Using the limiting velocity method, what is the permissible velocity (fps) for water flowing in this nonvegetated channel?

   A. 6.0
   B. 3.5
   C. 2.7
   D. 4.0

5. What is the value of Manning's n for this nonvegetated channel?

   A. .037
   B. .020
   C. .030
   D. .025

6. Using Mannings equation, $Vp = \frac{1.49}{n} R^{2/3} S^{1/2}$, the hydraulic radius of the channel is calculated to be 1.32 ft. The channel cross section area is found from $A = Q/V$ and is calculated to be 1.93 ft^2. The engineer then assumes that the channel depth should be approximately 1.3 feet. He also assumes that the bottom width, d, can be estimated from $A = bd$ where $b + 1.93/1.3$ or 1.48 ft. What should he do next?

   A. Add 20% to the depth value and the bottom width value to provide adequate freeboard in case of a heavy rainstorm.
   B. Check to see if his approximations for depth and bottom width are reasonable by using the relationship

   $$R = \frac{bd + zd^2}{b + 2d\sqrt{z^2 + 1}}$$

   C. Calculate the top width of the channel by using the relationship, $t = b + 2dz$.
   D. Calculate the wetted perimeter value for the channel using the relationship $2d\sqrt{z^2 + 1}$ to determine flow resistance.

---

*Items enclosed in brackets contain information in their stems necessary for the solution of problems contained in later items in that group of items.

Table 5 (continued)

7. What can be said about the engineer's estimates of the values for the depth and bottom width of the channel?

   A. Both values are a reasonable approximation of the true values.
   B. Neither value is a reasonable approximation of the true value.
   C. The width estimation based on assuming a rectangular cross section is only slightly in error
   D. The depth approximation is based upon assuming that R = d and is quite accurate for this channel.

8. What are the final values which are necessary for the depth, (D) bottom width (b), and top width (T) of the channel if it is to operate at the capacity given in the first part of this problem and under the soil and slope conditions specified? Include the necessary freeboard (ft.).

   A. D = 1.3, b = 1.5, T = 9.26
   B. D = 1.6, b = 1.8, T = 11.1
   C. D = 2.0, b = 7.0, T = 15.0
   D. D = 2.4, b = 7.0, T = 18.0

A parabolic channel is to be designed to carry 25 cfs of water on a 4% slope. Because the soil is easily eroded, the designer decides to vegetate the channel with fescue which is to be unmowed. Use this information to answer questions 9 - 11.

9. What is the maximum permissible velocity for water flowing through this channel (fps)?

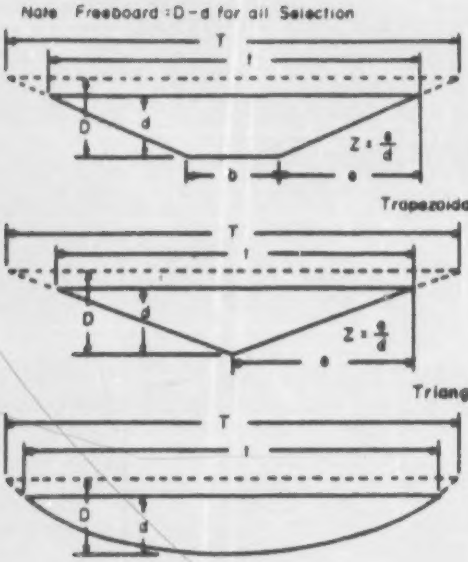
   A. 3
   B. 5
   C. 7
   D. 3.5

10. What is the retardance class for this vegetated channel?

   A. A
   B. B
   C. C
   D. D

11. What is the hydraulic radius of this channel?

   A. 1.1
   B. .58
   C. .82
   D. 1.6



	Cross-Sectional Area a	Wetted Perimeter, p	Hydraulic Radius $R = \frac{a}{p}$	Top Width
Trapezoidal Cross Section	$bd + Zd^2$	$b + 2d\sqrt{Z^2+1}$	$\frac{bd + Zd^2}{b + 2d\sqrt{Z^2+1}}$	$t = b + 2dZ$ $T = b + 2DZ$
Triangular Cross Section	Zd^2	$2d\sqrt{Z^2+1}$	$\frac{Zd}{2\sqrt{Z+1}}$ or $\frac{d}{2}$ approx.	$t = 2dZ$ $T = \frac{D}{d}t$
Parabolic Cross Section	$\frac{2}{3}td$	$t + \frac{8d^2}{3t}$	$\frac{t^2 d}{1.5t^2 + 4d^2}$ or $\frac{2d}{3}$ approx.	$t = \frac{a}{0.67d}$ $T = t\left(\frac{D}{d}\right)^{1/2}$

Figure 8. Properties of typical channels.

Table **10** Limiting Velocities and Tractive Forces for Open Channels. (Straight after Aging)

		For Clear Water		Water Transporting Colloidal Silts	
Material	n	Velocity, fps	Tractive Force, psf	Velocity, fps	Tractive Force, psf
Fine sand colloidal	0.020	1.50	0.027	2.50	0.075
Sandy loam noncolloidal	0.020	1.75	0.037	2.50	0.075
Silt loam noncolloidal	0.020	2.00	0.048	3.00	0.110
Alluvial silts noncolloidal	0.020	2.00	0.048	3.50	0.150
Ordinary firm loam	0.020	2.50	0.075	3.50	0.150
Volcanic ash	0.020	2.50	0.075	3.50	0.150
Stiff clay very colloidal	0.025	3.75	0.260	5.00	0.460
Alluvial silts colloidal	0.025	3.75	0.260	5.00	0.460
Shales and hardpans	0.025	6.00	0.670	6.00	0.670
Fine gravel	0.020	2.50	0.075	5.00	0.320
Graded loam to cobbles when noncolloidal	0.030	3.75	0.380	5.00	0.660
Graded silts to cobbles when colloidal	0.030	4.00	0.430	5.50	0.800
Coarse gravel noncolloidal	0.025	4.00	0.300	6.00	0.670
Cobbles and shingles	0.035	5.00	0.910	5.50	1.100

From Lane (1955).

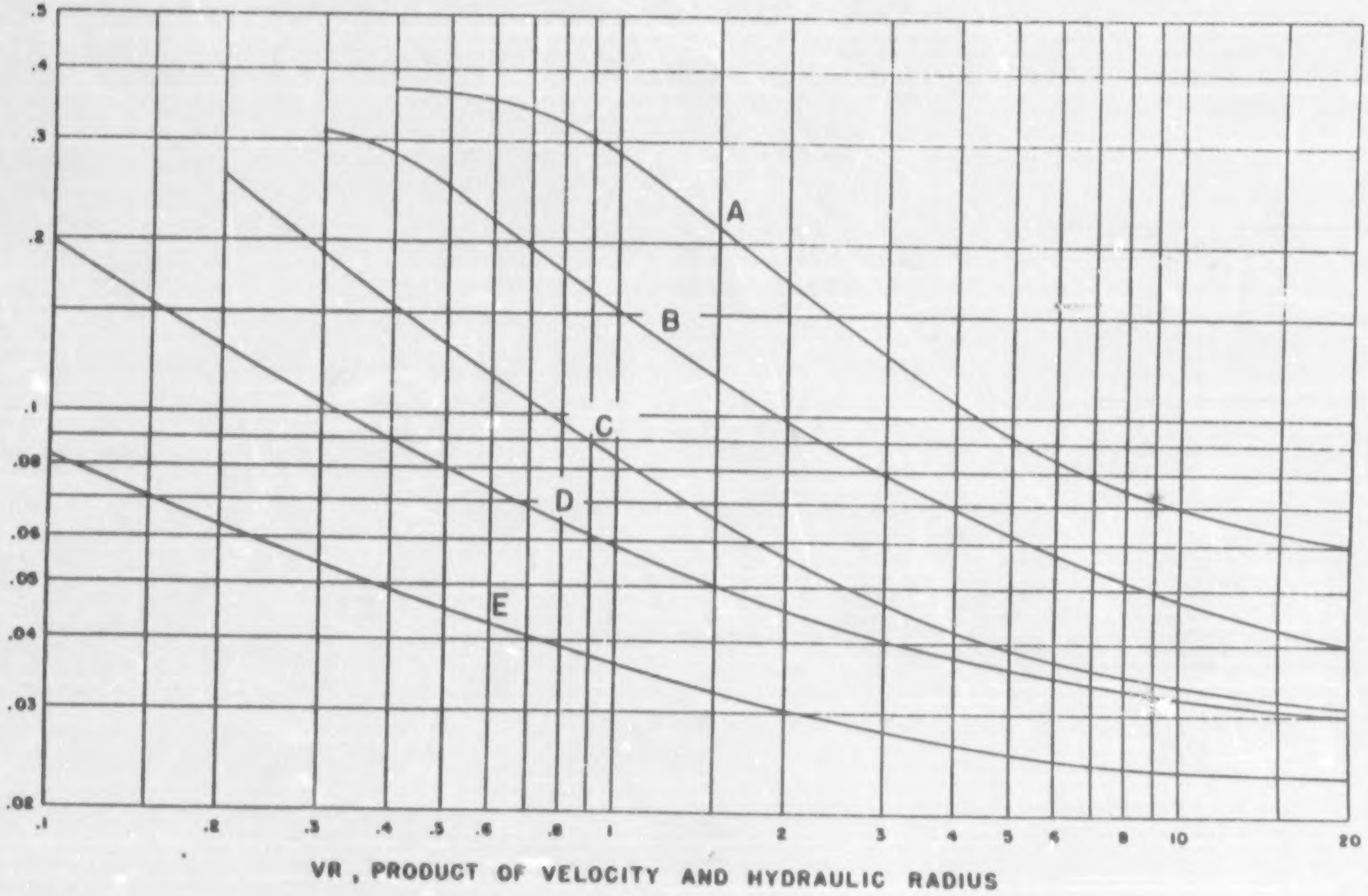
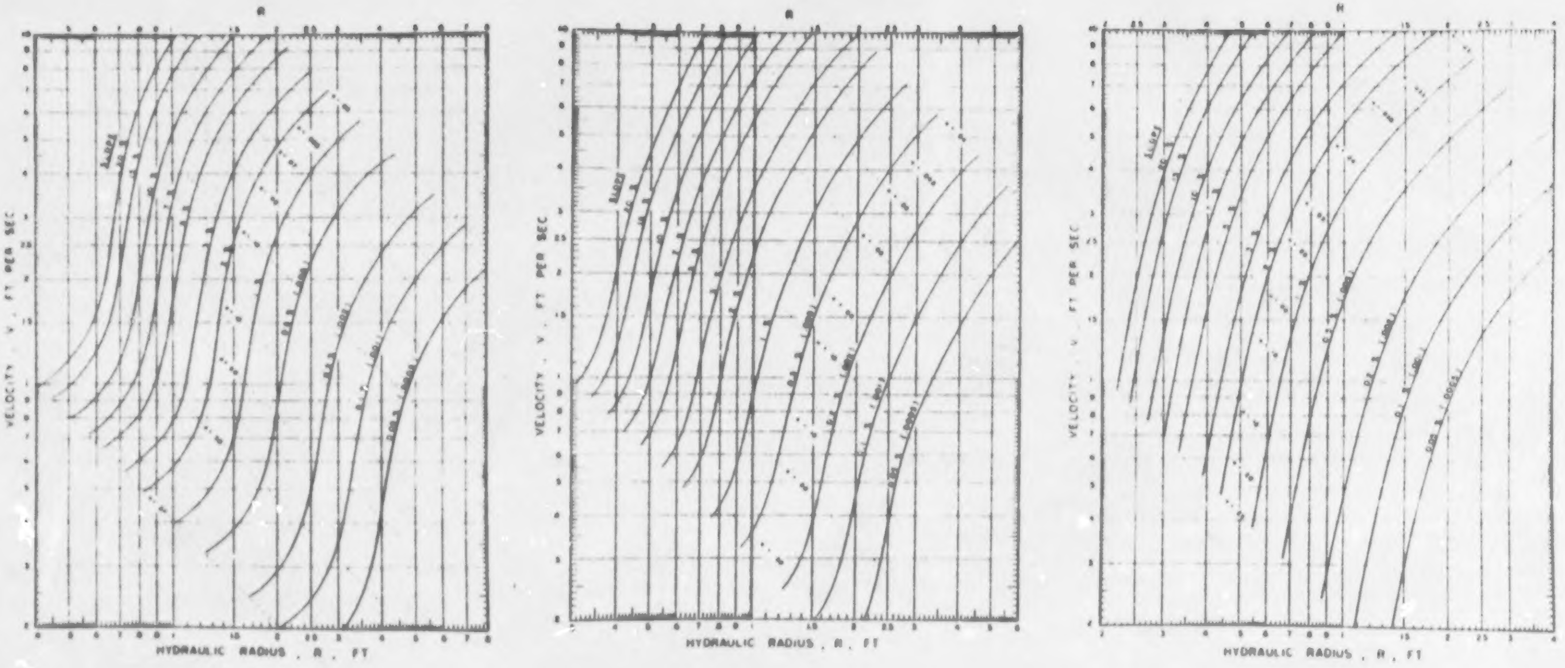


Figure 9. n-VR for various retardance classes.

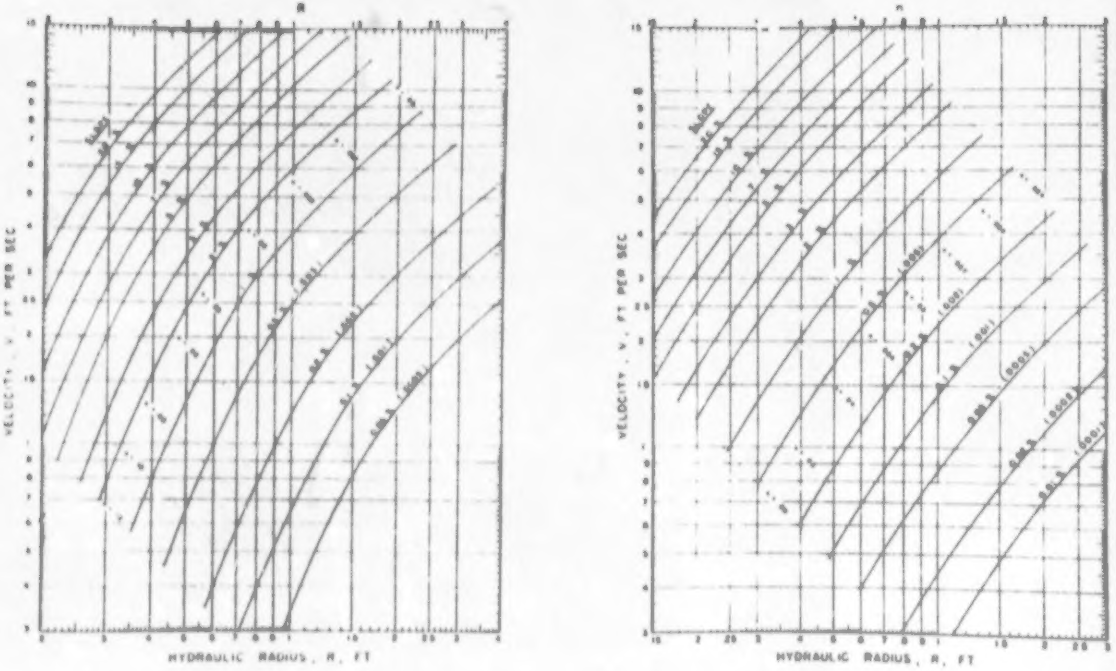


Retardance Class A

Retardance Class B

Retardance Class C

Figure 10. Solution for Manning's equation, vegetated waterways.
Retardance classes A, B, and C. (SCS, 1947)



Retardance Class D

Retardance Class E

Figure 11. Solution for Manning' equation, vegetated waterways. Retardance classes D and E. (SCS,1947).

Since this is an important part of what is taught in the course, and also of what is required in the real work setting, it is appropriate to require such tasks on the performance test.

Producing Items Which Test for Various Levels of Skill and Knowledge

For the most part the test items do not require computation. Rather they require knowledge of relationships and procedures. Items one through three test for knowledge of basic properties and relationships. It would be possible to develop a number of items to test for relationships other than those presented. Those items developed and included on the test ought to be central and important to wise use of the procedures being taught.

Items 4 through 8 represent a problem parallel to a practice problem given in the course for this unit. The example problems in each of the chapters in the manual define the functional competencies expected of students. These are the performance objectives for each unit or chapter. Therefore, it is best to prepare items which test for knowledge of the procedures and methods which are required for solution of the problems. Although the problem presented in items 4 through 8 is parallel to the practice problems used in the course, it presents different soil, slope, and other characteristics than were encountered in the practice problem. The new problem is parallel with respect to the skill and knowledge required for its solution, but not simply another identical problem where the individual need only substitute in new values to obtain the correct results. Each test item in the series of five items attempts to measure some particular aspect of the person's knowledge and skill in using the procedures to solve the problem. In addition, each item is written to be independent of the other items in the series with respect to having to have the correct answer to one item in order to have the correct answer to a later item. It is permissible to have a related series of items about a common set of problem situations, as long as the answer to any one item does not depend upon the answer to any other item.

Items 4 and 5 test for knowledge of how to enter the correct tables and extract the correct value for two variables given certain problem conditions. Item 6 measures the concepts related to proper estimation procedures in this type of problem. Item seven is a similar item. It tests for knowledge of when it is appropriate to apply a rule of thumb; the rule being that for shallow, wide channels, d is approximately equal to R.

Item 8 is the only item so far which requires any computation. It requires the individual to use information given in the original statement preceding item 4 and the additional information given in item 6. From this information the specifications for the channel can be calculated and the freeboard values determined. This is the most difficult and time consuming item. A few computational items of this type are needed in order to insure a wide range of item difficulties and to assess persons' knowledge and skill across the range of performance required for solving these types of complex problems. In other units of the course, such as the one dealing with the universal soil loss equation with all of its very complex four parts, it would be best not to require the working of a problem involving all of the parts of the equation. Rather, three of the values in the equation might be presented as already having been determined with the fourth to be determined from the appropriate use of information provided in a problem or question and through the selection and use of appropriate nomographs, rules of thumb, tables, and approximation procedures. Once again, items which test for knowledge of how to test the validity of the approximate solutions achieved by these procedures should be included, since this is an important intended outcome for the course.

Items 9, 10, and 11 are intended as another problem series which tests for knowledge of procedure, rules of thumb, and checking on estimation procedures for a channel of a different shape to be designed with vegetation. The complete question series through the checking on the estimation procedure values and the calculation of the final design specifications are not presented in the sample test.

However they could be developed the same way as is illustrated in the item 4 through 8 series. Again, because of time constraints only one or two of these computational items for each unit or chapter should be used with more of the other types of items which test for basic knowledge of concepts, relationships, and procedures.

When a series of related items and stimulus information at the beginning of these items is to be used by students, it is important to tell the persons being tested that series of items are presented in places; that the information given in the stem of the question and the other introductory information is needed in other questions; but that a wrong answer to any one question does not necessarily mean that all the remaining questions in the series will be incorrectly answered. It is also important to enclose any question series in a well defined bracket marked on the margins of the test item booklet, as indicated on the sample items.in order to indicate which items share common information in their stems.

Some General Guidelines

The general guidelines which follow may be helpful in designing test items for technical courses similar to the "Hydrology & Sedimentology" course.

1. Test for What has Actually Been Instructed -- Present only problems or questions which were actually instructed in the

short course itself. The Hydrology and Sedimentology text has much additional information and detail that must necessarily be omitted in the short course presentation. Participants cannot have been expected to have studied the text thoroughly by the end of the short course, but they can be expected to have understanding of basic procedures such as how to set up a problem; which models, assumptions, rules of thumb, and basic parameters to use; and how to extract desired values and graphic solutions to certain equations from charts, tables, and nomographs. This is what the test items should test for.

2. Be Sure Performance Objectives, Test Items, and Demonstration Problems are Congruent -- Use the structure of the actual problems used as instructional example problems as the operational description of what it is persons should be able to do at the end of the course. Those problems should be very clear, designed explicitly to illustrate the concepts and procedures to be learned, and can be broken down into individual test items to assess competence in each phase of the procedures in each section of the course.

3. Develop Test Items Which Map the Full Range of Performance -- Design several types of questions for each unit or chapter. These should be graded in difficulty from easy to difficult and should include:

A. Basic information and concept questions concerned with definitions, terms, and simple concepts upon which the other procedures depend. Item 2 on the sample test is such an item.

B. Basic relationship questions which test for comprehension of the relationships between physical variables and their representation in equations, graphs, etc. For example, a question about the point on the inflow and outflow hydrograph where the time of maximum storage occurs can test comprehension of such a relationship. Another question might be written to ask why the time of maximum storage is where the outflow and inflow hydrographs cross. Four answers could be provided with one correct and the other three being good distractors. Items 3 and 6 on the sample test assess this type of performance capability.

C. Procedural questions which test whether or not the person recognizes the proper steps in setting up a problem, working through the solution to a problem, etc. Notice the emphasis upon recognition. In such items the basics of the problem should be presented and several alternate ways of setting up the problem would then be given. Only one would be correct. The others would all be in error in some way, because of the misapplication or failure to adjust a model for a particular set of conditions, etc. This is what item 6 on the sample test is designed to do. The other items which require the person to recognize the correct values for

Manning's n or for maximum permissible flow given soil type and other information also do this. The person must recognize the correct value or be able to match those presented against those he or she looks up in a table or chart.

D. Concept application problems or questions which test the degree to which the person understands where a particular concept, rule of thumb, procedure, or method applies and does not apply. This type of item can often be written by giving the physical description of a problem and then having the student recognize the correctness of the approach or approaches outlined by which to solve the problem. An example of this type of item is number 7 on the sample test.

4. Prepare Brief and Time Efficient Test Items -- As is clear from the above discussion, most items should test for recognition of correctness of procedure, application of concepts, setting up of problems, and reasonable variable values as outcomes. Relatively little emphasis should be given to computational items because there is too little time to do so. In addition, any such test will be only a test of basic knowledge and skill in using the ideas and procedures in the manual, not in facility in actually applying the ideas and concepts in a highly accurate and wise manner. The latter outcome is desirable and can be achieved but not within the time limits of the short course. The

test items should be a reasonable sample of performance that can be expected to result from the actual short course instruction. A delayed post test or work samples of practicing engineers can be used to assess long term continued growth of knowledge and skills weeks or months after short course completion if so desired.

5. Provide Materials and Information Needed to Solve the Problems -- Information about formulas, equations, charts, tables of values, and nomographs should be provided. Test items should test for knowledge of how to use and apply such relationships, not for recall of formulas and relationships. These facts can be looked up by any practicing engineer and are routinely. The charts, formulas, graphs, tables, and nomographs needed to answer questions ought to be clustered together in sections for portions of the test to be available to persons, but also to provide a test of their ability to discriminate from among and properly use the appropriate equation, chart, or table. Example test items 4 - 6 attempt to illustrate this practice.

6. Use Standard Procedures to Produce Good Multiple Choice Items -- Follow the usual procedures for the design of good multiple choice items. A set of these general procedures is provided in a listing in Table 10. It should be apparent that multiple choice questions are very time efficient both from the standpoint of the time required for completion of the test and for scoring. However, any of the

items presented in the sample test could be used as an essay, constructed response, or problem solving item. The stems of good multiple choice items always have this property. If the stem is well designed it can be used in either the objective multiple choice format or as a constructed response item. Persons interested in the details of constructing multiple choice items may refer to Maratuza (1977) or other similar sources referenced in Chapters 10 and 11.

Table 10

LIST OF CONSIDERATIONS FOR PREPARING MULTIPLE CHOICE ITEMS

1. Is the item stem clearly written for the intended group of examinees?

2. Is the item stem free of irrelevant material? (Sometimes in a complex problem question you may want some irrelevant givens to test the person's knowledge of which relationships to use.) See sample item 1. "Parabolic cross section" could be deleted, but its presence requires some discrimination of irrelevant information from relevent information.

3. Is a problem clearly defined in the item stem?

4. Are the choices clearly written for the intended group of examinees?

5. Are the choices free of irrelevant material? (Again, the 3 false distractors need to be false and so may make use of given material which is normal to the problem but irrelevant to the given aspect being tested in a particular item.)

6. Is there a correct answer or a clearly best answer?

7. Have words like "always", "none", or "all" been removed from options?

8. Are likely examinee mistakes used to prepare incorrect answers for options?

9. Is "all of the above" avoided as a distractor?

10. Are the choices arranged in a logical sequence (if one exists)?

11. Was the correct answer randomly positioned among the available options?

12. Are all repetitious words or expressions removed from the choices and included in the item stem? (Example - item 4 on the sample should have fps in the stem, not after each distractor.)

13. Are all of the choices of approximately the same length? (Persons tend to select the longest option, and the longest option is also more often the correct one.)

Table 10 (continued)

14. Do the item stem and choices follow standard rules of punctuation and grammar?

15. Are all negatives underlined? (Example - which factor is not related to the numerical value for Manning's n?)

16. Are grammatical cues between the item stem and the choices, which might give the correct answer away, removed?

17. Is the item format appropriate for measuring the intended objective?

18. Are items independent from one another in terms of the answer to item n + 1 not being dependent on item n, etc.?

AUTHOR INDEX

Airasian, P. W., 201
Aleamoni, L. M., 7

Barfield, B. J., 22, 39, 51, 55 107, 114, 138, 141, 151, 166, 284, 308
Bellack, A. A., 5
Berliner, D. C., 146
Bloom, B. J., 58
Bloom, B. S., 6, 176, 177, 181, 201
Blythe, D. K., 22, 39, 141
Box, G. E. P., 68
Bramble, W. J., 95
Briggs, L. J., 56, 58, 165, 178
Bugelskik, B. R., 104, 112

Carroll, J. B., 58
Cleaver, T. G., 136, 208
Cohen, P. A., 136, 137, 251
Cole, H. P., 41, 147

Enell, J. W., 78, 264
Ericksen, S. C., 136

Ferry, R., 22, 39, 141

Gage, N. L., 146
Gagne, R. M., 7, 56, 58, 104 112, 137, 147, 165, 178
Greenfield, L. B., 6
Greenfest, M. C., 264
Grobman, H., 6, 7
Grogan, W. R., 137, 138, 208

Haan, C. T., 51, 55, 107, 114, 138, 151, 166, 284, 308
Hales, L. N., 165
Hambleton, R. K., 230, 252
Hastings, J. T., 6, 177, 181 201
Heimback, C. L., 136
Holland, J. L., 47
Hoyt, D. P., 244
Hunter, W. C., 68
Hunter, J. S., 68

Jones, J. A., 2, 31

Kliebard, H. M., 5
Klus, J. P., 2, 31
Knowles, M. S., 36
Kulik, C. C., 136, 137, 208, 251
Kulik, J. A., 136, 137, 208, 251

Lacefield, W. E., 93, 248, 259
Lavin, D. E., 239, 244
Livingston, S. A., 230, 250, 252
Lord, F. M., 61, 271, 307

Madaus, E. F., 6, 177, 181, 201
Manning, W. H., 58, 165, 179
Maratuza, V. E., 165, 201, 224, 233, 254
Marion, R., 178
Marshall, J. C., 165
Martin, E. D., 264
Mason, E. J., 95
McClelland, D. C., 239, 244, 296, 297
McCullough, R. C., 34
Mertens, D. M., 33, 42 46, 73
Miller, D. B., 42
Millman, J., 201, 226, 228
Morris, A. J., 46
Morstain, B. R., 42
Moss, P. J., 22, 39, 141

Novick, M. R., 61, 230, 252 271, 307
Nunnally, J. C., 165, 217, 233

Paradise, N. E., 112, 178

Salvendy, G., 178, 179
Sanders, J. R., 6, 21
Scriven, M., 28, 46
Seymour, W. D., 178, 179
Sherrill, P., 46
Shoemaker, D. M., 61, 271, 307
Smart, J. C., 42
Snelbecker, G. E., 112
Stice, J. E., 239, 244, 296

Thorndike, R. L., 165
Tyler, R. W., 5, 146, 165, 224, 247, 252, 294, 297

Webb, W. B., 56, 165
Wiesehuegel, R. E., 42, 188
Wolf, R. M., 22, 165, 294
Work, C., 237, 238, 248
Worthen, B. R., 6, 21